Social Media Operationalized for GIS: The Prequel

Full Paper

Anthony J. Corso
California Baptist University
acorso@calbaptist.edu

Abdulkareem Alsudais
Claremont Graduate University
abdulkareem.alsudais@cgu.edu

Abstract

With social media a de facto global communication channel used to disseminate news, entertainment, and one's self-revelations, the latter contains double-talk, peculiar insight, and contextual observation about real-world events. The primary objective is to propose a novel pipeline to classify a tweet as either "useful" or "not useful" by using widely-accepted Natural Language Processing (NLP) techniques, and measure the effect of such method based on the change in performance of a Geographical Information System (GIS) artifact. A 1,000 tweet sample is manually tagged and compared to an innovative social media grammar applied by a rule-based social media NLP pipeline. Evaluation underpins answering, prior to content analysis of a tweet, does a method exist to support identifying a tweet as "useful" for subsequent processing? Indeed, "useful" tweet identification via NLP returned precision of 0.9256, recall of 0.6590, and F-measure of 0.7699; consequently GIS social media processing increased 0.2194 over baseline.

Keywords

Social Media Grammar, NLP, GIS, Sharing Economy.

Introduction

Accurate and timely dissemination of information is an imperative function to support a population fixated on news, entertainment, and extreme consumerism. Sparse text social media being used as a proxy for formal information dissemination outlets is fundamentally unexplored. However, social media such as Twitter, Reddit, or FaceBook, provide real-time data for both the corporate and consumer population, and each are certainly connected. For example, the increasing use of Twitter as a real-time communication channel escalates the need to explore tweets that not only propagate user sentiment but also contribute to formal matters such as business strategy creation, decision making in critical situations, and regulation (Quattrone et al. 2016). Natural Language Processing (NLP) of a tweet corpus will better facilitate an understanding of social behavior, reveal connections between disparate geographic groups, or enhance analysis in information system artifacts. Several researchers have provided theoretical insight and practical application of social media analysis, social media parts-of-speech (POS) tagging, and social media spatial context analysis (Bramsen et al. 2011; Hiruta et al. 2012; Stefanidis et al. 2013).

Although social media NLP pipelines exist and facilitate promising results, only elementary solutions, if any, grammatically tag sparse text social media (e.g., text content of a tweet) and identify it as "useful" or "not useful." Furthermore, Geographical Information Systems (GIS) artifacts import various data types and provide robust qualitative and quantitative analysis tools making them a stable platform with the capability to consume operationalized social media (Corso and Alsudais 2015). A supervisory framework is integral for processing, applying an English-like grammar, and GIS consumption of social media. Such a framework would better allow scholars, practitioners, bureaucratic agencies, and sharing economy entities to increase their capabilities with efficient and effective use of sparse text social media, allocate and manage human resources, promote community awareness, deploy resources, etc. Authenticating such claims results in substantive theory and empirical examination of research, construction and evaluation of a formative grammar-tagging social media NLP pipeline, and an innovative approach to answer the question, prior to content analysis of a tweet, does a method exist to support identifying a tweet as "useful" for subsequent processing in a predictive crime analysis social media GIS artifact? A favorable

answer represents a best practice framework for processing tweets in order to reach equilibrium between tweets that are "useful" and tweets that are "not useful," such that, the former can be further processed and the latter expunged. Additionally, such a processing step expands GIS investigative characteristics into orthogonal problem domains previously not considered.

The aptitude to implement a social media grammar in order to harness knowledge from the abundance of sparse social media would bring state-of-the-art capability to many organizations. Although distinct disciplines deal with prediction of social behavior, customer behavior, business resource deployment, public safety announcements, or Big Data research, many are searching for additional variables that do not infringe on known attributes but conceivably extend important foresight to solve their particular problem. An interrelated issue addresses processing, analyzing, and interpreting sparse text social media's Big Data characteristics. On one hand, social media data are streaming, in many instances in real-time, and there are vast amounts of it to collect, process, and analyze. On the other hand, sparse data content (e.g., a tweet) can be authored such that in many, many cases it is not at all useful. Accurately and confidently separating "useful" from "not useful" tweets, and training a classifier via Big Data criteria, exhibits great difficulty and is extremely expensive in terms of both human and machine resources. Therefore, to improve status quo and solve problems the aforementioned disciplines are experiencing, further research and implementation of germane solutions are needed.

The primary objective of this paper is to propose a novel pipeline to classify a tweet as either "useful" or "not useful" by using widely-accepted NLP techniques; then, measure the effect of such method based on the change in performance of a GIS artifact. The novel concept of unlocking sparse text social media's latent linguistic features provides a revolutionary framework to explore variables orthogonal to social media in a meaningful way. Unequivocally, the artifact constructed consumes a Twitter corpus and overcomes data sparsity of a tweet to discover linguistic-based latent grammatical structures embedded within. All research conducted is supported by proven fundamental linguistic experiment, NLP analysis, and common information retrieval measures and information system controlled experiment inquiry (Leroy 2011). The solution exemplifies the value of real-time data collection, social media linguistic processing, and Big Data assimilation with respect to intelligence-based GIS artifact construction. Consequently, the project explicitly acknowledges the possibility of a relationship between a grammatically processed social media tweet corpus and "useful" or "not useful" tweets as the resultant, which can be subsequently imported into a predictive crime-based social media GIS artifact, to support various grouping analysis constructs. Beyond the introduction is the literature review and a section describing the problem and solution developed. The experiment section describes hypothesis development and the research methodology including data and its features. An analysis section describes the inquiry, quantitative features, and results. Last, a conclusion and suggestion for future extension section is provided.

Literature Review

Research of sparse text social media classification explores the NLP continuum from simple key-word event extraction to parsing the frequency of obscure bigrams used for identification of structural relationships between words (Jurafsky and Martin 2009). One element of sparse text classification rests on tokenization, i.e., unit of measure for a block of text-based content. Well-known researchers offer pronounced vision into the theoretical, practical, and evaluation of this discipline (Frantzi et al. 2000; Hirst and Feiguina 2007; Phan et al. 2008; Sriram et al. 2010; Piao and Whittle 2011). While many of their topics are not of great recent interest, other communities will benefit by focusing on their foundation of unstructured social media corpora analysis. The work of Hirst and Feiguina (2007), Phan et al. (2008), Sriram et al. (2010), and Piao and Whittle (2011) have many research traits in common. Each builds robust learning models by identifying NLP features of common domain artifacts where they are used in classification of a short text corpus. All of them calculate a baseline for their respective project's corpus based on what feature best lends itself to the linguistic component of unit analysis and easy comparison to like units. That is, all four use subject matter expert analysis and additional machine-based statistical analysis to conduct evaluation of their model. Piao and Whittle (2011) are instrumental because their work extends Frantzi et al. (2000) in calculating automatic recognition of n-grams based on the c-value metric and develop capability to use an evaluation metric based on linguistic and statistical analysis. A number of novel classifiers that are given sparse text inputs exist; but in at least one conclusion subpar

results are indeed based on such sparse text inputs (Phan et al. 2008). Within this research thread the consensus for working with short texts is that text sparseness makes processing difficult within a natural language processing pipeline (Hirst and Feiguina 2007); a key issue this work overcomes.

Several researchers address periphery issues with respect to geographical information systems, Big Data, and social behavior when sparse text corpora are coalesced with natural language processing in a GIS solution. In terms of Big Data's veracity characteristic, isolation of tweet contextualization as a key issue or requirement for tweet summarization is deemed significant (Zingla et al. 2015; Zubiaga and Ji 2014; Torres-Moreno 2014). Each represent the issue with a slightly different methodology, e.g., association rules mining (Zingla et al. 2015), analysis by professional (Zubiaga and Ji 2014), and statistical algorithms (Torres-Moreno 2014). Overall, each solution is novel and leads users to a more credible Twitter-based social media experience. Evaluation of information system tweet development methodologies with automatic tweet tagging to experiment with the social behavior side of tweet research is also considered (Alonso et al. 2013; André et al. 2012; Hurlock and Wilson 2011). In particular, the issues addressed are the captive nature of the tweet and the subjective way it is interesting (Alonso et al. 2013), worthy of reading (André et al. 2012), or usefulness (Hurlock and Wilson 2011). Each use a traditional research model and measure by having subject matter experts participate in the study. With the aid of human interaction they compare machine versus human judgment of tweets (Alonso et al. 2013), derive value from user ratings of tweets (André et al. 2012), and implement a traditional user study with results supporting "useful" or "not useful" tweets as a component of grounded theory analysis (Hurlock and Wilson 2011). None implement a social media NLP pipeline to distinguish grammatical structure.

Last, a research area where the spatial and temporal presence of crime is fused with social media. In such an environment a personification of one's life is actively explored, and geospatial crime mapping using tweet context analysis can indeed enhance accuracy of predicting specific types of crime. Bendler et al. (2014) use a normalized Big Data tweet corpus. Gerber (2014) validates a similar risk-based artifact using a decision support system. Last, Corso et al. (2016) extend the social risk model prudently offered by Drawve (2014), Kennedy et al. (2011), and Caplan et al. (2011) to solidify the foundation of social media being used as a proxy for social behavior. Since it is assumed that social media, in a similar way to traditional GIS crime risk layers, can be operationalized, and subsequently compared via the same performance priorities of traditional layers. Collectively, literature only makes a preliminary investigation and analysis of sparse text social media's scope of use on relevant GIS artifacts.

Problem Description

In this section, the need to apply an NLP pipeline to reduce the size of a tweet corpus is discussed. Also, the rationale behind the inquiry into the development of a classifier that labels tweets as either "useful" or "not useful" for additional examination is explained. Last, implications of GIS artifact performance when developing such a classifier are argued.

Crafting a natural language processing pipeline for a social media corpus with very, very limited English grammar and data structure manifests many challenges. For example, a tweet corpus exhibits message content of 140 characters, contains a sizeable amount of noise, and needs significant data preprocessing to determine meaning, is an exemplar. Within this purview a novel yet observable solution needs to decipher a tweet's latent grammatical structure and make an assessment about the usefulness of its textual content. The urgency for developing such an artifact is growing, especially if one considers a near real-time social media information dissemination paradigm. A vast amount of literature documents how a tweet's data sparsity and unique dialect, e.g., acronym, acrostic, emoji, and emoticon use, must be overcome before a text mining classification pipeline is applied; not an easy problem to address and at best difficult to solve. Presently, contextualized, interesting, or noteworthy tweets are identified only after content sparseness and English grammatical structure are interpreted. However, it is conceivable that a sparse content, unique dialect, short-text-processing application distinguishing "useful" from "not useful" tweets, prior to message context examination, would be extremely innovative and revolutionary.

The definition of a "useful" tweet is a tweet that contains meaningful text content; thus, should be subject to subsequent contextual analysis. Whereas, tweets that are "not useful" are ones that do not contain meaningful text content and thus should be deleted because further processing will not reveal any relevant information. A sample of five tweets selected from these two categories is depicted in Table 1. A

structured and reproducible process to select potentially useful tweets from a corpus, before message context is identified, originates for many reasons. For example, a GIS artifact supporting the analysis of social media as a proxy for social behavior (consumer perception of advertising, geographical event detection dissemination, consumer sentiment, geographic dialogue analysis, etc.) would be required to evaluate contextual meaning of each tweet in a corpus, even if their textual content was meaningless and not useful to the solution. Furthermore, better utilization of GIS analytics are achieved since the artifact does not waste processing cycles to extricate absolutely useless tweets. Creating a corpus of useful tweets to be subsequently consumed by a GIS is extremely efficient, effective, and overall good practice.

Useful	Not Useful
I'm at Old Navy in Chicago IL https://t.co/lczpuR9NLF	ballloooonnsss ??? http://t.co/mjhuKyH6DM
My Phone Die So Fast	WAYYOHANDSIDETOOSIDEPUTEMINTHEAIR
David Bowie is... my favorite! @ David Bowie is At Mca Chicago	Funniest
I aint gta stunt on nobody...trust me Yo LoL!	#CuppyCoffee!!!!!!!???
I can't wait to see lora tomorrow	@_TylorShane ????

Table 1. Sample of Useful and Not Useful Tweets

The magnitude of such an issue is realized when processing commenced on a social media tweet corpus that contains approximately 415,000,000 tweets; it was collected between September 1, 2014 and April 21, 2016. Attempts to process the entire dataset in a GIS artifact proved problematic, and additional steps are required to reduce the size of the corpus. With the original long-term objective to develop a social media predictive crime analysis GIS artifact, randomly removing tweets from the corpus will compromise it by potentially deleting significant crime-related tweets. Accordingly, it is deemed required to develop an NLP pipeline to process each tweet and classify it as either "useful" or "not useful" for additional examination. Social media corpora processing a posteriori approximates 60 percent of tweets are useful for ensuing GIS analysis. With a bona fide 415,000,000 tweets needing to be processed, approximately 138,000,000 would be labeled "not useful" and simply discarded. A successful implementation of a pipeline will reduce the size of the corpus while safeguarding relevant and potentially significant tweets.

In addition, the social media predictive crime analysis GIS artifact needs the tweet corpus to be grammatically tagged. Conceivably, by using one of the many parts-of-speech (or tagsets), which include a description (noun, verb, pronoun, preposition, adverb, etc.) of each word (or token) in the English corpora tagging process. Approaches developed include the Stanford tagger, Penn Treebank part-of-speech tagset, 87-tag Brown corpus tagset, and various custom tagsets developed for a specific corpus. However, traditional tagsets produce inadequate results when applied to a social media corpus, whereas custom tags produce better results, significant adjustments are necessary. Moreover, both traditional and custom tagsets exhibit the same issue, i.e., they are not developed to tag the unique characteristics of a social media corpus. This is a significant issue because a predictive crime analysis GIS artifact needs to consume variables that are used as a proxy for a traditional risk-based crime analysis. Therefore, the social media corpus must represent social behavior and have the ability to associate with other behavior layers in the construction of a social media predictive GIS crime analysis artifact, e.g., Supplemental Nutrition Assistance Program (SNAP) locations. SNAP is a nutrition assistance social program offered to low-income participants were SNAP benefits are provided to low-income individuals and families so they can access eligible food items using Electronic Benefits Transfer (EBT) cards across the nationwide network of locations. Crime is also a variable associated with social behavior and therefore the third component used in the procedure. Collectively, social behavior is modeled by the variables consumed by the GIS and every tweet collected needs to be tagged and stored even if it is ultimately deemed incomprehensible for further linguistic processing. A better more efficient way exists to identify tweets that exhibit the possibility to be subsequently processed. Thus, a small-scale linguistic tagging preprocessing procedure needs to be developed.

Methodology

In this section, the proposed method for processing, operationalizing, and labeling either "useful" or "not useful" geotagged tweets is discussed. An initial data cleaning process that facilitates subsequent NLP

processing and GIS data consumption, is also performed. The cleaning process detects unexpected processing errors and damaged data. Additionally, the code checks for latitude, longitude, and text field content; it also verifies one tweet per line is properly saved to secondary storage. Whereas each tweet consists of latitude, longitude, and content fields (used explicitly in this inquiry) thirty additional attributes and their data also need processing. Therefore, best practice suggests the GIS import metadata, attributes, and content only for contextually useful tweets and discard the rest.

As a result, the General Architecture for Text Engineering (GATE) natural language processing suite of tools was selected. The GATE NLP application constructed executed an NLP pipeline that had three functions and is shown in Figure 1. First, each tweet's text field needs to be tokenized. Tokenization in the pipeline was set to break on space, views the text content of each tweet, and identifies each group of text characters with its own token id number. Second, each token needs to be tagged with an appropriate part-of-speech tag; Figure 2 denotes text, tokens, and part-of-speech tags of a sample tweet.

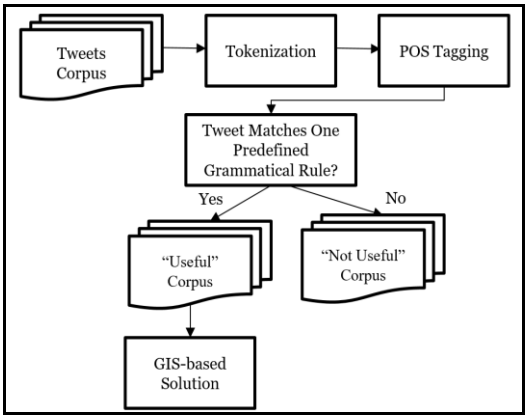

Figure 1. Proposed NLP Pipeline

We in Chicago wit the Homies https://t.co/TQbw

```
We in Chicago wit the Homies https://t.co/TQbWI
{category=PRP, kind=word, length=2, orth=upperInitial, string=We}
{category=IN, kind=word, length=2, orth=lowercase, string=in}
{category=NNP, kind=word, length=7, orth=upperInitial, string=Chicago}
{category=IN, kind=word, length=3, origString=wit, orth=lowercase, string=with}
{category=DT, kind=word, length=3, orth=lowercase, string=the}
{category=NNS, kind=word, length=6, orth=upperInitial, string=Homies}
{category=UH, kind=URL, length=23, replaced=9, rule=URL, string=https://t.co/TQbWhBa
{category=UH, kind=punctuation, length=1, position=endpunct, string="}
{category=SYM, kind=symbol, length=1, string=<}
{category=:, kind=punctuation, length=1, string=-, subkind=dashpunct}
{category=:, kind=punctuation, length=1, string=-, subkind=dashpunct}
{category=:, kind=punctuation, length=1, string=-, subkind=dashpunct}
{category=NNP, kind=word, length=3, origString=NYC, orth=allCaps, string=NYC}
```

Figure 2. Tweet Text, Tokens, and, POS Tags

The last part of the NLP pipeline was to apply a list of social media grammar identification rules. Table 2 shows the grammatical rules implemented by the NLP application and loaded into its pipeline. In this step the identification of the grammatical structure of a tweet—to be further processed by a particular solution—was implemented. In other words, an individual tweet was deemed to be "useful" or "not useful" if one or more grammar rules were applied to it.

Tag	Rule
"Tweet-SymbolUID-Text"	{Token.category==NN}|{Token.category==JJ}|{Token.category== NNP}| {Token.category==NNS}|{Token.category==INS}
"Tweet-SymbolUID-TagHTTP"	{Token.category==NN, Token.kind==word, Token.length==4, Token.orth==lowercase, Token.string==http}
"Tweet-SymbolUIDTagAndTokenTagVerb-PRP"	{SymbolUIDTagAndTokenTagVerb}{Token.category == PRP}| {SymbolUIDTagAndTokenTagVerb}{Token.category == "PRP$"}| {SymbolUIDTagAndTokenTagVerb}{Token.category == POS}
"Tweet-SymbolUIDTagAndTokenTagNoun-PRP"	{SymbolUIDTagAndTokenTagNoun}{Token.category == PRP}| {SymbolUIDTagAndTokenTagNoun}{Token.category == "PRP$"} {SymbolUIDTagAndTokenTagNoun}{Token.category == POS})

Table 2. NLP Tags and Rules

In total, 29 rules that yielded twelve custom tags were created. A sample of four rules and four tags are presented in Table 2. Tweets tagged with one or more rules were marked useful, and stored for additional GIS processing, and untagged tweets were deleted.

Three measures were computed to evaluate the performance of the method. The results of these measures are in Table 3 which is located the results section below. Once the "useful" tweet corpus was generated, the next step was to process the corpus in the GIS artifact. The custom developed pipeline would be deemed successful if the GIS solution performed significantly better when the "operationalized" corpus was used.

The geographic information system constructed for the project had two key functions; exploratory analysis via visualization of latent structures. Two, statistical analysis of impact of social media as a proxy for social behavior. When combined with other social behavior variables such as supplemental nutrition assistant program locations, used as a proxy for social behavior, and crime locations, the independent variable structure of the GIS analytic artifact was determined. Collectively, the GIS artifact; that is, tweet locations, SNAP locations, and crime locations, respectively.

Experiment

The main objective is to conduct a randomized controlled experiment to compare the impact of "useful" from "not useful" tweets on GIS processing. As such, testing the variations between "useful" tweets, SNAP locations, and crime as social behavior predictive crime GIS layers via simplistic GIS analytic grouping analysis is useful. With the use of random selection, within tweet, SNAP, and crime layers used for analytic GIS group processing, there were three treatments with one condition being the baseline, one condition being expert analysis (two judges subjectively labeling the tweets), and one condition being the NLP pipeline. This predicts, the more "useful" the tweet corpus the better the GIS artifact will support the general research question, further proffer Figure 3 as a supervisory framework processing infrastructure, and identify a workflow for this work and similar inquiry.

Hypothesis Development. Based on preceding arguments, greater GIS grouping analysis accuracy revolves around a three-step social media solution (capture, process, and tag) a tweet corpus. Where the tagging solution implemented an NLP pipeline supporting custom parts-of-speech tags that represent a proxy for social behavior. To investigate such an outcome and help readers better understand the landscape, consider the problem statement, the NLP pipeline in Figure 1, and Figure 3 (noted below, and as it reveals the supervisory framework in its entirety); thus, Figure 4 and Hypothesis 1 are offered:

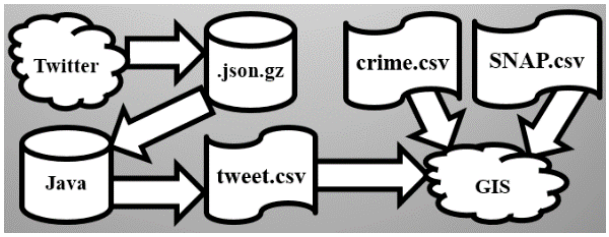

Figure 3. Research Framework

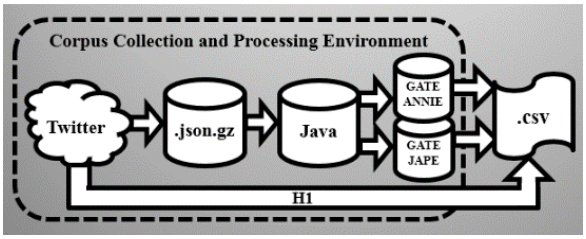

Figure 4. Hypothesis 1

Hypothesis 1: A Geographical Information System will have higher social behavior predictive capability when processing an NLP-tagged tweet corpus rather than a standard-tagged or expert-tagged tweet corpus.

The hypothesis will focus on the independent variable the GIS artifact with three distinct treatments and the dependent variable of GIS grouping analysis variance. In other words, greater precision of parts-of-speech as applied to short text social media will create a social media grammar structure used to better group social behavior layers. As a result, multidimensional aspects of the research model will capture many latent social facets of a tweet corpus. In turn, GIS grouping analysis R-squared value will increase.

Independent Variable. This is the tweet corpus, SNAP, and crime data consumed by the GIS. Collectively, it is a particular version of the social media predictive crime analysis GIS artifact. The treatments applied to the corpus and the control of the other social behavior variables. The independent variable of an improved social media predictive crime analysis GIS artifact will be the result. The improved process represents the result of how much better the custom NLP pipeline tagged a tweet corpus. In effect, the task reflects how well the pipeline was able to match tweet grammar rules applied by human subject matter experts or random chance.

Dependent Variable. Total or overall variation of the GIS grouping analysis results will be used for the dependent variable and is selected because its calculation is deemed to be significantly influenced by the improved or degraded treatment applied to each GIS condition.

Sampling Corpus. The general descriptive statistics used for demonstration purposes in this work considered 1,000 randomly selected tweets. Subsequently, using the NLP pipeline and breaking on space, the corpus was tokenized into 1,748 sentences and 20,512 tokens. The sentence tokens were reconfigured to match the original tweet, i.e., one tweet equal to one sentence, and assigned to the baseline, expert, and pipeline condition. The baseline treatment is to act as the control group without any treatment.

Evaluation Plan. The evaluation component of the project is achieved in two parts. Evaluation of the pipeline and evaluation of the GIS processing solution. First, after preprocessing and removal of damaged data, evaluation unit of measure was at the individual tweet level. That is, each tweet's content had the tokenization, tagging, and rule-based treatment applied. One treatment is the baseline the other a modified NLP pipeline. Then, both the baseline case and the NLP grammar treatment case allow precision, recall, and F-measure to be calculated. The relationship between the social media tweet variable as operationalized in this study's model will be examined via GIS exploratory analysis. Subsequent to tweet processing and evaluation, analysis was visualized in informative ways, i.e., statistical measures provided for a rigorous evaluation of attribute relationships as grouped via GIS solution. More specifically, the GIS grouping analysis tool was configured to identify the relationship between the social media corpus and other social attributes. The grouping analysis input features class was the joined locations layer of tweet, SNAP, and crime. Explanatory variables were based on count and proximity.

Results

NLP Procedure

The objective of this task was to evaluate the performance of the NLP pipeline that labeled tweets as either "useful" or "not useful." To accomplish this, 1,000 tweets were randomly selected from the corpus. These tweets were then manually annotated by two judges and tagged as either "useful" or "not useful." After preprocessing and applying the rule-based NLP pipeline process for the 1,000 tweets, 20,512 tokens were retrieved. The method was evaluated by comparing labels generated using the NLP pipeline to ones generated by the judges. In addition, 1,748 sentence tokens were reconstructed to represent the original integrity of each tweet. No remaining or unidentified tokens were represented; thus, the entire 1,000 tweets attempting to be identified by conducting the experiment were tagged. Either the tweet received one of twelve custom tags or simply one of the many standard part-of-speech tags used in the standard GATE Hepple Tagger. The baseline values are based on randomly labeling half of the tweets in the dataset as "useful" and half as "not useful." The Baseline and Treatment measures for precision, recall, and F-measure are noted in Table 3. Baseline precision was 0.7022 and Treatment precision was 0.9256. Comparison of the proposed artifact to a baseline is a large component of evaluation; with calculations being based on the approach of one tweet per line of sentence content token. Through simple observation of the proposed new approach better results were evident. Theoretically, the evaluation provided better overall results because of the pipeline. Thus, the given calculations in Table 3, and the overall comparison for the solution being based on precision, recall, and F-measure, the new artifact proposed yields approximately a 0.3181 gain in a post and pre artifact precision measure.

	Precision	Recall	F-Score
Baseline	$\frac{500}{712}$	$\frac{500}{1{,}000}$	$\frac{2 \times .7022 \times .500}{.7022 + .500}$
	0.7022	0.5000	0.5841
Treatment	$\frac{659}{712}$	$\frac{659}{1{,}000}$	$\frac{2 \times .9256 \times .6590}{.9256 + .6590}$
	0.9256	0.6590	0.7699

Table 3: Metrics for Baseline and NLP Treatment

GIS Procedure

The purpose of this task was to measure the impact of the proposed pipeline on the performance of a GIS solution. The proposed method is deemed satisfactory only if the performance of the GIS solution

increases when an operationalized tweet corpus is used. The performance of the GIS solution was evaluated when three input sets were used: 1) A baseline set that consist of randomly selected tweets that were not processed so that tweets that are "not useful" are removed, 2) An expert sets that comprised of tweets that two judges deemed "useful," and 3) A "proposed pipeline" set that consist of tweets that were tagged as "useful" by the proposed NLP pipeline.

The map in Figures 5, 6, and 7 is a street map of downtown Chicago, and fishnet spaced at 750 feet with tweet, SNAP, and crime locations displayed. When all attributes are combined and exploratory analysis of the latent underlying structures are represented via grouping analysis, various cell shading results. Figure 5 represents GIS grouping analysis results for the baseline condition. Table 4 displays the R-squared value for each social behavior variable given, i.e., Baseline, Expert, and Proposed Pipeline treatments. In Table 4, the Tweet attribute for the Baseline solution preformed as expected, i.e., 0.5146 percent. In the Proposed Pipeline, the Tweet attribute contributed to the overall solution posting an R-squared value of 0.6275. Overall the Proposed Pipeline provided for an increase of .2194 percent increase over baseline, which suggests that the employing the NLP pipeline improved the performance of the GIS artifact.

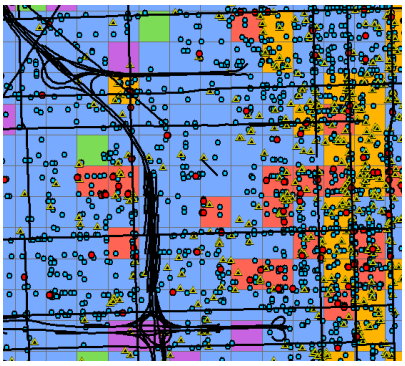

Figure 5. Baseline

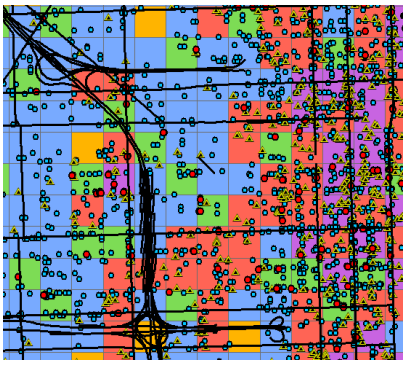

Figure 6. Expert

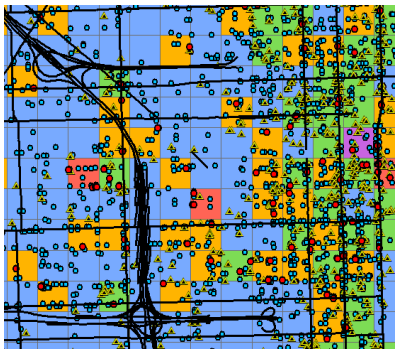

Figure 7. Proposed Pipeline

	Tweet	SNAP	Crime
Baseline: R-squared	0.5146	0.6556	0.8945
Expert: R-squared	0.7155	0.6320	0.7571
Proposed Pipeline: R-squared	0.6275	0.7126	0.7396

Table 4: Metrics for Baseline, Expert, and NLP Treatment

Discussion

Research in the area of tweet content analysis suggests it is associated with three specific, yet closely interrelated, dimensions of artifact design. The association between these dimensions—social media normalization, grammatical tagging, and usefulness—are examined to uncover quantifiable results and perhaps meaningful, qualitative, and visualized relationships often required in both data analysis and information dissemination artifacts. The innovative social media grammar applied to the tweet corpus was used to produce sophisticated class labels that were subsequently compared to traditionally used English grammar tweet corpus tags. Although the key proxies for evaluation are expert level a further more robust solution is possible. Within this theoretical framework and the experiment presented, the research model supports an artifact that is evaluated and scrutinized via quantitative method.

Limitations

Limitations of the work include not only corpus collection and tokenization but also notable potential for error. With currently more than half a billion tweets a day being sent via Twitter's infrastructure the former is addressed given the concept of random selection. The tweet corpus was collected within a latitude- and longitude-specific polygon and therefore the issue becomes whether or not the corpus

collected was a representative sample given this enormity. Perhaps the so called sophistication of the user plays a component role in non-randomness of collection given that users actively participation in being located. There are two perspectives on the issue; one, API used to collect the tweets from the Twitter stream; two, time and location of tweet selection. The collection process used the standard REST API therefore the overall percent of tweets collected off the stream was very small. Conversely, other projects, e.g., the corpus collected by Stanford University via Twitter's firehose API collect a sizable sample from the Twitter stream. The second limitation arrives from the simple break on space tokenization method applied to the corpus. More refined tokenization methods should be used to enhance tokenization. This perhaps tainted results given the nature of sparse tweets or how the default POS tagger assigned the tag.

Conclusion and Future Work

This inquiry investigated the application of an innovative NLP pipeline to evaluate its role in the outcome of predicting "useful" or "not useful" tweets. It also illustrated the need for such an application via implementation of a GIS artifact implementing social behavior attributes. It commenced because opportunity to fill the gap between sparse text social media part-of-speech tagging and application of a context free grammar suggesting a latent grammar exists; such that it can help determine if a tweet is useful for successive processing. It concludes, for the first time, that grammatical knowledge of such a sparse text social media corpus promptly provides inquiry of meaningful content while purging not useful structures. Also, a grammar-based social media suggests opportunity to identify less researched, yet intriguing, counterpart of sparse or acronym-based short message social media being used as a news, entertainment, or real-time critical event distribution source. By means of chance, expert analysis, and a rule-based classifier this project hypothesized a link between sparse social media, an appropriate NLP treatment, and a social media grammar where the solution was tested and verified. The contribution is innovative, and suggests advanced development of grammar-based sparse text social media corpora NLP treatments are needed; they can support a positive and more meaningful artifact.

Future work will attempt to advance the research area with greater detail being considered via NLP capability that applies a higher degree of grammatical tagging of the tweet corpus. Such that tweets are tagged in granular approach as compared to a chance-based baseline or expert tagged social media corpus. This concept tests the relationship between social media and grammar when controlling for levels of grammatical tagging. A favorable outcome produces an artifact with better capability to predict useful or comparative tweets based on how grammatically correct it is. The work is simply a foundation for subsequent social media processing, e.g., implementation of social media as an orthogonal variable in a sharing economy GIS artifact.

Another critical area for future research will expand the reach of social media's real-time communication network via integration with predictive geographic information system solutions. In particular, the specific area to be addressed is social behavior risk modeling. In this predictive analysis discipline a gap exists because social media risk modeling deals with social behavior variables, i.e., variables that are a byproduct of social risk factors—bars, night clubs, strip clubs—tend to fit the category. More specifically, social media might be orthogonal to risk and be used as a behavioral risk indicator or proxy risk source. For example, population or income are not social behavior risk factors but may be used as orthogonal variables to associate risk and subsequent possibility of social behavior. Ensuing NLP tweet corpus research need the ability to incite positive predictors and will need to better status quo to explain variations in the dependent variable.

REFERENCES

Alonso, O., Marshall, C.C., and Najork, M. 2013. "Are Some Tweets More Interesting Than Others?# Hardquestion," *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*: ACM, p. 2.

André, P., Bernstein, M., and Luther, K. 2012. "Who Gives a Tweet?: Evaluating Microblog Content Value," *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*: ACM, pp. 471-474.

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

Bramsen, P., Escobar-Molano, M., Patel, A., and Alonso, R. 2011. "Extracting Social Power Relationships from Natural Language," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 773-782.

Caplan, J.M., Kennedy, L.W., and Miller, J. 2011. "Risk Terrain Modeling: Brokering Criminological Theory and Gis Methods for Crime Forecasting," *Justice Quarterly* (28:2), 2011/04/01, pp. 360-381.

Corso, A., Alsudais, K., and Hilton, B. 2016. "Big Social Data and Gis: Visualize Predictive Crime," *AMCIS Conference 2016 Proceedings*).

Corso, A.J., and 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*), pp. 1-29.

Frantzi, K., Ananiadou, S., and Mima, H. 2000. "Automatic Recognition of Multi-Word Terms:. The C-Value/Nc-Value Method," *International Journal on Digital Libraries* (3:2), pp. 115-130.

Gerber, M.S. 2014. "Predicting Crime Using Twitter and Kernel Density Estimation," *Decision Support Systems* (61), pp. 115-125.

Hirst, G., and Feiguina, O.g. 2007. "Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts," *Literary and Linguistic Computing* (22:4), pp. 405-417.

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

Hurlock, J., and Wilson, M.L. 2011. "Searching Twitter: Separating the Tweet from the Chaff," *ICWSM*, pp. 161-168.

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

Kennedy, L.W., Caplan, J.M., and 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), pp. 339-362.

Leroy, G. 2011. *Designing User Studies in Informatics*. Springer Science & Business Media.

Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. 2008. "Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections," *Proceedings of the 17th international conference on World Wide Web*: ACM, pp. 91-100.

Piao, S., and Whittle, J. 2011. "A Feasibility Study on Extracting Twitter Users' Interests Using Nlp Tools for Serendipitous Connections," *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*: IEEE, pp. 910-915.

Quattrone, G., Proserpio, D., Quercia, D., Capra, L., and Musolesi, M. 2016. "Who Benefits from the "Sharing" Economy of Airbnb?," in: *Proceedings of the 25th International Conference on World Wide Web*. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, pp. 1385-1394.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. 2010. "Short Text Classification in Twitter to Improve Information Filtering," in: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. Geneva, Switzerland: ACM, pp. 841-842.

Stefanidis, A., Crooks, A., and Radzikowski, J. 2013. "Harvesting Ambient Geospatial Information from Social Media Feeds," *GeoJournal* (78:2), 2013/04/01, pp. 319-338.

Torres-Moreno, J.-M. 2014. "Three Statistical Summarizers at Clef-Inex 2013 Tweet Contextualization Track," *CLEF (Working Notes)*, pp. 565-573.

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

Zubiaga, A., and Ji, H. 2014. "Tweet, but Verify: Epistemic Study of Information Verification on Twitter," *Social Network Analysis and Mining* (4:1), pp. 1-12.