Aug 10th, 12:00 AM

Will the Home Team Win? On the Road to 1.5 Billion Tweets and Six Thousand Baseball Games Providing Insight!!!

Anthony Corso
California Baptist University, acorso@calbaptist.edu

Nathan A. Corso
California Baptist University, nathan.corso@calbaptist.edu

Will the Home Team Win? On the Road to 1.5 Billion Tweets and Six Thousand Baseball Games Providing Insight!!!

Emergent Research Forum (ERF)

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

Nathan A. Corso
California Baptist University
nathan.corso@calbaptist.edu

Abstract

Researchers operate with limited budgets and inadequate resources. This prohibits big data research and suppresses innovation needed to direct inquiry and construct robust research-based information systems. Such issues are not insuperable, e.g., this project is initialized with limited resources and attempts to build theory, describe architecture, and set the vision for future work. This first "On the Road to …" paper tenders a methodology that examines the use of social media variables as a proxy for human emotion and epistemic activity. A social media corpus is processed and a regression model considers MLB team wins as the dependent variable and a social media tweet corpus, operationalized via NLP, as the independent variable. Results are presented. Future work describes a predictive GIS artifact that will input, process, and visualize a spatial and time-based, NLP processed, social media corpus and is integration with geospatial indexing.

Keywords

Big GIS Data, Geospatial Indexing, Social Media, Natural Language Processing.

Introduction

Powered by ease of data collection and the readiness of authors to microscopically share every aspect of their diurnal observations and emotional state, big data social media analysis is becoming integrated with a firm's strategic goals (Hanna et al., 2011). This new era of business is sparking colossal demand for research platforms that collect, analyze, and visualize data (Haugh & Watkins, 2016). Geographic Information Systems (GIS) and their aptitude for analytic and visualization capabilities are a critical link in the discussion. Comingling social media and GIS analysis is gaining popularity and can successfully strengthen myriad predictive information systems (McKitrick et al., 2022) (Ristea et al., 2020). Sports Fans Social Media Research commenced in the new millennium to drive business revenue (Hur et al., 2007) and over time is being adjusted to account for both fan observations (Fernández-Gavilanes et al., 2019) and fan emotion within the course of a sporting event (Yu & Wang, 2015). Beyond driving revenue, detecting and highlighting sports fans' in-game observations and emotion is relevant for media analysis, comradery of a city's citizens, or a sports team owner (perhaps identifying a latent competitive home team advantage). To date, a limited, yet very focused, number of research projects investigate the detection and aggregation of Sports Fans Social Media Emotion (SFSME) (Agrawal et al., 2018). A number of reasons exist that make the creation and automation of methods to detect SFSME complex. First, the processing power necessary for data collection, storage, analysis, and visualization. Second, the subtle nature of text analysis techniques applied to the data (commonly called a corpus). In particular, Natural Language Processing (NLP) feature engineering is challenging in general but is extremely difficult for the sparse text of a social media corpus. Third, an end-to-end research platform where results are visualized in a near-real-time meaningful way.

Current research is engulfed with collecting social media and identifying significant features. However, simultaneously describing a particular event and identifying an emotional state in a methodologically advanced, viable, valid, and valuable way (Toivonen et al., 2019)—specifically given a big data corpus—is extremely uncommon. A GIS and its capacity for data management, analytic capabilities, and system architecture solves the problem. It is the intended artifact and this work investigates a framework to

preprocess a big data social media corpus for GIS consumption. In addition, an evaluative regression model of social media, including empirical and emotional game factor variables, is proposed. Similar models show the potential to achieve performance increases in the analysis of the emotional state of an event, e.g., tweet-level sentiment detection (Yue et al., 2019). Furthermore, resources utilized for this project are very limited. Corpus size is extremely small, independent variable count is limited to a few, and overall large-scale visualization is nonexistent. Contributions of this paper are trifold; first, identify the main challenges in and provide a valid framework for big data social media GIS research and analysis given limited human and financial resources. Second, via NLP feature engineering, validate fundamental linguistic structures extracted from a social media corpus that are successively consumed by a GIS. Third, introduce a framework that results in useful visualizations for stakeholders. In addition, the work supports GIS research beyond retrospective outcomes, i.e., intends to spark the interest of researchers wishing to delve into epistemic social media variables being used as a proxy to measure the state of human emotion and their potential for predictive capabilities (Corso & Alsudais, 2017). Subsequent sections consist of a literature review, methodology section describing data and its features, regression structure, and hypothesis. Results and discussion sections follow. Conclusions and recommendations for future research complete the document.

Related Work

Social media demonstrates the ability to support intelligence-based analysis provided solid methodological approaches are used to convert data to knowledge (Castillo et al., 2021). Latent features of social media corpora can be explored, e.g., via Natural Language Processing (NLP) (Yue et al., 2019) (Corso & Alsudais, 2017) and various other techniques (Volkova et al., 2015). Sundaram et al. (2013) provide insight on relational, object-oriented, and similar database technologies that address large-scale query and performance solutions with respect to searching social media. Logical progression integrates social media with GIS artifacts and is of particular interest with respect to predictive spatial temporal analysis and visualization. A model by Jia et al (2016) set precedent for interactive tweet corpus analytics with a processing scale of greater than one billion tweets. Their solution introduced an analytics and visualization system called Cloudberry. In terms of Su et al. (2018), a Cloudberry model was extended with a middleware solution to query a tweet corpus of over a billion twitter feeds streaming in real-time. More recently the Zika virus was explored via the Cloudberry architecture to investigate Twitter data of large scale Masri et al. (2019). Such methodologies provide support for GIS artifacts integrated with geospatial social media. Consequently, social media's potential contribution to predictive GIS artifact construction is significant; thus, new visualization and spatial integration technologies need exploration. GIS visualization techniques are cartographically distinct yet social media layer addition is easily achieved. Geospatial indexing algorithms are noteworthy and greatly enhance social media's use in GIS artifacts. Several geospatial grid systems exist, e.g., Figure 1 displays the S2, GeoHash, Hilbert, and H3 indexing systems. GIS social media indexing research literature is scant; hence, future direction of literature for social media GIS integration needs to focus on geospatial indexing.

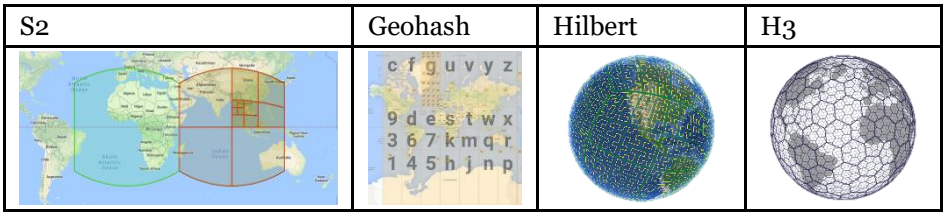

Figure 1. Geospatial Indexing

Methodology

Corpus

Primary data collection of a Twitter tweet corpus occurred between September 20, 2014 and February 28, 2022. Overall, seventy data collection sites were identified—as displayed in Figure 2, thirty (30) Red Pins, one for each major league baseball stadium, thirty-two (32) Blue Pins, one for each professional NFL football stadium, and seven (7) Shaded Rectangles, college-saturated geographic locations. Data collection

for each site was accomplished by applying a latitude and longitude polygon bounding box in the collection code. For example, each location considers a Southwest (bottom left) corner pair of coordinates in the form (33.137051, -112.511466) and a Northeast (top right) corner coordinate pair in the form (33.767319, -111.531636). The collection code implements a Java library with the JSON metadata collection format. The corpus size from all sites exceeds 3.5 billion tweets, of which, 1.5 billion tweets are from the 30 MLB sites.

Figure 2. Data Collection Sites

Natural Language Processing

The Python-based Natural Language Toolkit (NLTK) library is used to construct a tweet annotation pipeline (Bird et al., 2009). Individual tweets are the unit of measure; specifically, from each tweet JSON object the "text" element's content is extracted. Table 1 displays example content and identifies the features to be parsed from each tweet. Sentiment values are 1 for happy, -1 for sad, and 0 for undetected. Polarity values are 1 for positive, -1 for negative, and 0 for neutral. Cognition values are 1 for present and 0 for undetected.

	Tweet Text Element Content	Sentiment	Polarity	Cognition
Tweet One	Lol	1	1	0
Tweet Two	☺ we Won	1	1	1
Tweet Three	nothR loossseeee	-1	-1	1
Tweet Four	Great day at fenaaawayyy Sox Won	0	1	1
Tweet Five	@<name> have a safe trip down and enjoy the game! Looking forward to your coverage.	1	1	1

Table 1. Tweet Text and NLP Features

Regression Setup

Linear regression is useful for analysis since one dependent variable is possibly influenced by one or more independent variables. Furthermore, P-values can be calculated for each variable; as such, the following regression equation is given: Population = $\beta_0 + \beta_1 * X_{i1} + \beta_2 * X_2 + \beta_3 * X_3 + \epsilon_i$. In terms of the social media features identified in Table 1 the regression equation is as follows:

Win = intercept + slope 1 * "polarity" + slope 2 * "sentiment" + slope 3 * "cognition" + error

Experiment and Results

Building on NLP social media feature engineering, parsing tweets, and fundamental concepts of regression, it is expected that predictive capabilities of an information system artifact yield better results than chance. It is asked; what effect does home team social media vigor have on home team win percentage? On one hand, equal chance refers to 50% probability of the home team wining. On the other hand, greater home team social media vigor suggests more wins. The NLP feature extraction process of identifying polarity, sentiment, and cognition represent vigor. Subsequently, if vigor is infused with other GIS feature classes, spatial temporal analysis can be scrutinized in quantitative and qualitative ways. Such a theoretical

framework tests the relationship between SFSME and team wins while controlling for multiple levels of emotion and epistemic activity; a hypothesis is considered.

Hypothesis Ho (Null). There is no relationship between the number of home team vigorous social media posts and home a team win.

Hypothesis H1 (Alternative). There is a positive correlation between the number of home team vigorous social media posts and a home team win.

Dependent Variable. The variable representing the outcome of a process; herewith, a win.

Independent Variables. The three variables (sentiment, polarity, and cognition) representing vigor. More specifically, variables developed via NLP techniques applied to each sampling unit, i.e., a single tweet; where, each may systematically influence the dependent variable.

Nuisance Variables. Potential nuisance variables are still under consideration.

Random Assignment. Assignment of sampling units is conducted via the default assignment of tweet ID as pulled from the collection process.

The regression model was executed for two MLB teams, Chicago Cubs and Chicago White Sox. For each team, home game date and outcome data was obtained. Total number of games is displayed via Observations row in Table 2 and Table 3. Based on game date, tweets were selected and NLP processed to construct an individual corpus for each team. The experiment was conducted in a usability lab at California Baptist University and results are displayed in Table 2 and 3.

Chicago Cubs			*Coefficients*	*P-value*
Multiple R	0.201453481	Intercept	1.242102367	0.274265108
R Square	0.040583505	hapSENcubs	-0.104971215	0.388962189
Adjusted R Square	-0.066018328	sadSENcubs	-0.105543012	0.396327972
Standard Error	0.501695896	neuSENcubs	0.106239105	0.38532391
Observations	81	posPOLcubs	-2.40927E-05	0.890472995
Significance F	0.927567336	neuPOLcubs	-0.001194182	0.218361822
		negPOLcubs	-3.37528E-05	0.789221129
		yesCOGcubs	0.001983453	0.836343071
		noCOGcubs	3.334E-05	0.759539026

Table 2. Regression Results Cubs

Chicago White Sox			*Coefficients*	*P-value*
Multiple R	0.48249462	Intercept	0.661353805	0.285232784
R Square	0.232801059	hapSENsox	-0.04036742	0.46704672
Adjusted R Square	0.141195215	sadSENsox	-0.036632442	0.522980268
Standard Error	0.464981582	neuSENsox	0.040058351	0.469512594
Observations	76	posPOLsox	7.26977E-05	0.635220806
Significance F	0.017625714	neuPOLsox	-0.000830029	0.360895352
		negPOLsox	-0.00026446	0.04282572
		yesCOGsox	0.033424864	0.000898234
		noCOGsox	0.000127086	0.267131705

Table 3. Regression Results Sox

Discussion and Limitations

It was expected that results for each team would be similar. There are no significant independent variables for the Cubs. However, there are multiple significant variables for the Sox, i.e., 'yesCOGsox' (presence of cognition) and 'negPOLsox' (negative polarity). This bifurcation, given the similarity of corpora is enough to investigate the Significance F (P-Value) value for each team. For the Cubs a value of 0.928 is not significant; the Sox post a value of 0.0176, which, is significant. Thus, in the case of the Cubs, accept the Null and in the case of the Sox reject the Null hypothesis.

Limitations of the work fit into a few specific areas. First, sample size and random selection. With currently more than a billion tweets per day the issue of a representative sample is present. Second, the NLP procedure used to process a tweet places significant influence on the feature being extracted. This is both positive and negative, for example, the process could identify more, or less, of a content that does or does not exist. Third, the representation of independent variables as to the outcome of a game. Factors involved in game outcomes are manifold. Various other independent variables should be considered in order to draw strict conclusions.

Conclusion

Revolutionary geographic information system artifact construction implementing social media's latent features to explore the relationship with a dependent variable is possible. Tweets identified with social media vigor, i.e., polarity, sentiment, and cogitation provide mixed outcomes via regression analysis. In the

Chicago White Sox case, with p-values for cognition and positive polarity of 0.000898 and 0.0428, respectively, more investigation is needed.

NLP tokenization, beyond, break on space, social media part-of-speech tagging, and social media feature engineering are sophisticated techniques. Each must be properly crafted and evaluated with respect to corpus consumption via GIS. For example, each token in a tweet was assigned a part-of-speech tag via the NLTK universal tagset; why not use the Oxford or Brown tagset. Comingling GIS feature classes with social media via geospatial indexing will allow for fine grained analysis and needs substantial examination.

References and Citations

Agrawal, A., Gupta, A., & Yousaf, A. (2018). Like it but do not comment: Manipulating the engagement of sports fans in social media. *International Journal of Sport Management and Marketing*, *18*(4), 18.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Castillo, A., University of Granada, Spain, Benitez, J., EDHEC Business School, France, Liorens, J., University of Granada, Spain, Braojos, J., & Ramon Llull University / University of Granada, Spain. (2021). Impact of Social Media on the Firm's Knowledge Exploration and Knowledge Exploitation: The Role of Business Analytics Talent. *Journal of the Association for Information Systems*, *22*(5), 1472–1508. https://doi.org/10.17705/1jais.00700

Corso, A., & Alsudais, A. (2017). *Social Media Operationalized for GIS: The Prequel*.

Fernández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., & Javier González-Castaño, F. (2019). Differentiating users by language and location estimation in sentiment analisys of informal text during major public events. *Expert Systems with Applications*, *117*, 15–28. https://doi.org/10.1016/j.eswa.2018.09.007

Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. *Business Horizons*, *54*(3), 265–273. https://doi.org/10.1016/j.bushor.2011.01.007

Haugh, B. R., & Watkins, B. (2016). Tag Me, Tweet Me if You Want to Reach Me: An Investigation Into How Sports Fans Use Social Media. *International Journal of Sport Communication*, *9*(3), 278–293.

Hur, Y., Ko, Y. J., & Valacich, J. (2007). Motivation and Concerns for Online Sport Consumption. *Journal of Sport Management*, *21*(4), 521–539. https://doi.org/10.1123/jsm.21.4.521

Jia, J., Li, C., Zhang, X., Li, C., Carey, M. J., & su, S. (2016). Towards interactive analytics and visualization on one billion tweets. *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 1–4. https://doi.org/10.1145/2996913.2996923

Masri, S., Jia, J., Li, C., Zhou, G., Lee, M.-C., Yan, G., & Wu, J. (2019). Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic. *BMC Public Health*, *19*(1), 761. https://doi.org/10.1186/s12889-019-7103-8

Ristea, A., Al Boni, M., Resch, B., Gerber, M. S., & Leitner, M. (2020). Spatial crime distribution and prediction for sporting events using social media. *International Journal of Geographical Information Science*, *34*(9), 1708–1739. https://doi.org/10.1080/13658816.2020.1719495

Su, S., An, M., Perry, V., Jia, J., Kim, T., Chen, T.-Y., & Li, C. (2018). Visually Analyzing A Billion Tweets: An Application for Collaborative Visual Analytics on Large High-Resolution Display. *2018 IEEE International Conference on Big Data (Big Data)*, 3597–3606. https://doi.org/10.1109/BigData.2018.8622183

Sundaram, N., Turmukhametova, A., Satish, N., Mostak, T., Indyk, P., Madden, S., & Dubey, P. (2013). Streaming similarity search over one billion tweets using parallel locality-sensitive hashing. *Proceedings of the VLDB Endowment*, *6*(14), 1930–1941. https://doi.org/10.14778/2556549.2556574

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., & Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, *233*, 298–315. https://doi.org/10.1016/j.biocon.2019.01.023

Volkova, S., Bachrach, Y., Armstrong, M., & Sharma, V. (2015). *Inferring Latent User Properties from Texts Published in Social Media*. 2.

Yu, Y., & Wang, X. (2015). World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets. *Computers in Human Behavior*, *48*, 392–400. https://doi.org/10.1016/j.chb.2015.01.075

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, *60*(2), 617–663. https://doi.org/10.1007/s10115-018-1236-4