Blending Quantitative, Qualitative, Geospatial, and Temporal Data: Progressing Towards the Next Generation of Human Social Analytics

C.J. Hutto
Human Systems Engineering
Georgia Tech Research Institute (GTRI)
Atlanta, GA, USA
e-mail: cjhutto@gatech.edu

***Abstract*—Human social analytics in the next generation will need to embrace more multifaceted representations of human behavior with more complex models. Such models will need to integrate data of disparate forms, using disparate units of measure, collected from disparate sources, at disparate scales. Next generation social scientists will also face issues related to developing methods and tools to help facilitate the collection, processing, analyzing, and visualizing of such multifaceted social data. This paper illustrates these challenges by reporting on the development of a complex model of societal well-being (an inherently qualitative construct) which blends large scale quantitative, geospatial, and temporally referenced data of disparate forms, units, sources, and scales. We then demonstrate tools and methods intended to facilitate the progression towards next generational social analytics at large scales. We conclude by discussing several open questions with regards to social analytics, including those related to ethics and privacy concerns.**

Keywords-human centered data science; human social analytics.

I. INTRODUCTION

All sorts of human social and behavioral data are now available, and on unprecedented scales. Of course, social scientists still rely heavily on traditional sources of social and behavioral data such as in-person, telephone, or computer assisted interviews, questionnaires and survey instruments, and sources of "thick descriptions" [1] of human behavior compiled from ethnographic or anthropological observation research. However, new sources of human social behavior data are now available due to our increased use of mobile phone, GPS technology, and personal wearable technology (such as fitness trackers), as well as the digital traces of technology-mediated communications and online social interactions. These new data sources will allow researchers to conduct human social analytics for extraordinary levels of insights ranging from intra-individual scale investigations, through inter-personal and group level interactions, to organizational and even population scale research. Over the next 25 years (a generally accepted duration of a generation), social scientists and data analysts will need to modernize their ways of thinking about and interacting with human behavior data, else risk their research becoming obsolete and irrelevant.

In this paper, we address issues facing the next generation of social data scientists. We do so in the first part of the paper by presenting an example in which we progress beyond simple representations of human social behavior by constructing a complex model of individual and societal well-being. We describe the integration and analysis of data of varying forms, collected via diverse methods from a variety of sources by different groups, consisting of varied units of measure, spanning a temporal range of more than 40 years, and representing human behavioral data at disparate scales. In short, we present a case study of blending quantitative, geospatial, and temporally diverse data for the purpose of advancing human social analysis for an inherently qualitative construct using a more complex (and, we argue, more representative) model of human social behavior.

In the second half of this paper, we describe how new methods borrowed from the field of computer science can be leveraged to support next generation human social analysis of qualitative data. Computational natural language processing (NLP) and statistical machine learning (ML) techniques have the potential to be extremely useful for blending *thick data* (which is most commonly qualitative in form: e.g., descriptive text, audio, imagery, video, or similar multimedia) with the concepts of *big data* (typically more quantitative in nature). Here, we discuss three specific "tools" that embody NLP and ML techniques to support large-scale human social analysis on qualitative data. The first tool, called VADER (**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner), provides researchers the ability to quantify both the direction (positive or negative) and magnitude of affective expressions in textual documents ranging from word-level to tome-level scales, processing millions of sentences in a matter of seconds [2]. The second tool, CASTR (**C**ommon-ground **A**cquisition for **S**ocial **T**opic **R**ecognition), produces supporting text-based information needed to establish so called *common ground*, whereby sharing mutual facts and knowledge generally facilitates faster, better understanding [3], [4]. The third tool, EAGLE-ID (**E**thnicity, **A**ge, **G**ender, **L**iteracy/**E**ducation **Id**entifier), automatically aids in characterizing demographic features of individuals based on social profile data. Finally, we discuss how digital crowdsourcing economies such as Amazon Mechanical Turk (a massive, distributed, anonymous crowd of individuals willing to perform human-intelligence micro-tasks for micro-payments) can be leveraged as a valuable resource for the next generation of social science research and practice [5].

We conclude by discussing several open questions with regards to human social analytics, including those related to ethics, data ownership and use, and personal privacy concerns.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-519-7

II. INCREASING REPRESENTATIONAL COMPLEXITY OF DATA MODELS FOR HUMAN SOCIAL ANALYTICS

Traditional social scientific models of human behavior are often over-simplified representations of what in actuality are very complex aspects of the world. Human social analytics in the next generation will need to embrace more multifaceted representations of human behavior with more complex models. Such models will need to integrate data of disparate forms, using disparate units of measure, collected from disparate sources, at disparate scales. In this section, we contribute an example in which we develop a complex, system-of-systems representation of societal well-being.

A. From Simple to Complex Modeling of Well-being

Individual and societal constructs of well-being are well established in traditional social science and economic literature as a person's assessment of their own general *happiness* and overall *satisfaction* with their personal life [6], [7]. Following from [8], we further posit that happiness and satisfaction are themselves complex social constructs which holistically comprise four principal constituents:

1. Affective Experiences: the longer-term experiences of pleasant affect (as well as a lack of unpleasant affect) as indicated, for example, via their general perceived happiness in life, in their marriage, and with their cohabitation companion (e.g., partner or roommates).
2. Global Life Judgements: a person's overall belief regarding how interesting they find their own life in general (e.g., whether they consider life to be dull, routine, or exciting), as well as a judgement about the general nature of humanity (whether they believe most other people to be trustworthy, fair, and helpful).
3. Cognitive Appraisals: a person's subjective self-assessment of their own current socioeconomic state relative to their life goals, as well as broader social comparisons. Determinants include financial status self-appraisals, social status self-appraisals (e.g., social rank and social class), and self-appraisals regarding their health, the relative quality of their domicile, and aspects of the city in which they reside.
4. Domain Specific Satisfaction: the degree of fulfillment or contentment with important social elements such as satisfaction with their family life, friendships, hobbies and recreational interests, job/career, and their wages.

Traditional social analytics tend to focus on a narrowly scoped subset of the above constituents. While such studies do provide useful insights, they are limited precisely because they are narrow; due to the inherent interconnectedness of these constituents, complex interactions abound. Nevertheless, they hold much greater analytical value when they are considered in conjunction with one another. The whole is greater than the sum of its parts, and aggregate-level insights may never emerge unless and until the underlying relationships are expressly represented.

To this end, we present an example in which we incorporate 130 different manifest indicators for- and correlates of- individual and societal well-being. To do so, we blend qualitative, quantitative, geospatial, and temporal data from several sources. While detailed model specification is beyond the scope of this paper, we find the model useful as a reference for discussing next generation social analytics.

B. Blending Qualitative, Quantitative, Geospatial, & Temporal Data

The data for our complex model of well-being are drawn from several public data sets comprising records from 30 different collection activities spanning 42 years (from 1972 to 2014) across nine different divisions of the United States Census Bureau [9]. This data integrates 25 manifest indicators of societal well-being, organized into latent variable constructs representing the four principal constituents described in Section II-A. An additional 17 indicators provide data providing more objective measures of individual *quality of life and standard of living*, such as highest education level attained, number of people living in a household, type of dwelling (and whether owned or rented), various employment characteristics (part time, full time, student/homemaker, unemployed, retired, etc.), and constant (i.e., annual inflation adjusted) income in dollars. Also included are data capturing information about each respondent's *demographic* details, the *general political climate* (public opinion regarding amount of taxes paid, the efficacy of the courts, and national programs related to healthcare, transportation, and public transit), established local and regional *geographic boundary data*, annually recorded data regarding the *general economic climate* of the nation (such as inflation rates, consumer price indices, prime lending rates, and annual gross domestic product (GDP) per capital growth), and data characterizing the *general security climate* (e.g., individual and community exposure to crimes, perceptions of fear, etc.).

As one might imagine, the data are operationalized in multifaceted ways, taking multiple forms, units, and scales of measurement. In all, we integrate data from nearly 60,000 respondents spanning 42 years with regard to 130 different variables of interest, where each variable puts (on average) potentially 7 unique degrees of positive or negative pressure on individual and/or societal well-being. All told, this leverages approximately 55 million data points for our model, allowing for a very rich and complex representation of well-being – much more sophisticated than many other typical, prevailing social science models.

We argue that this representation, as opposed to a simpler model (for example, one based primarily on measures of *happiness*) is a more accurate reflection of true societal well-being. To illustrate this point, consider Fig. 1, in which we visually depict how a simplistic representation of well-being (happiness scales) compare to a more complex representation of societal well-being for different geographic regions in the United States. Different insights emerge (especially in the southern regions) when affective experiences, global life judgements, cognitive appraisals, domain specific satisfaction, objective socioeconomic quality of life and standard of living data, the general political climate, general economic climate, and the general security climate are incorporated when considering societal well-being.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-519-7

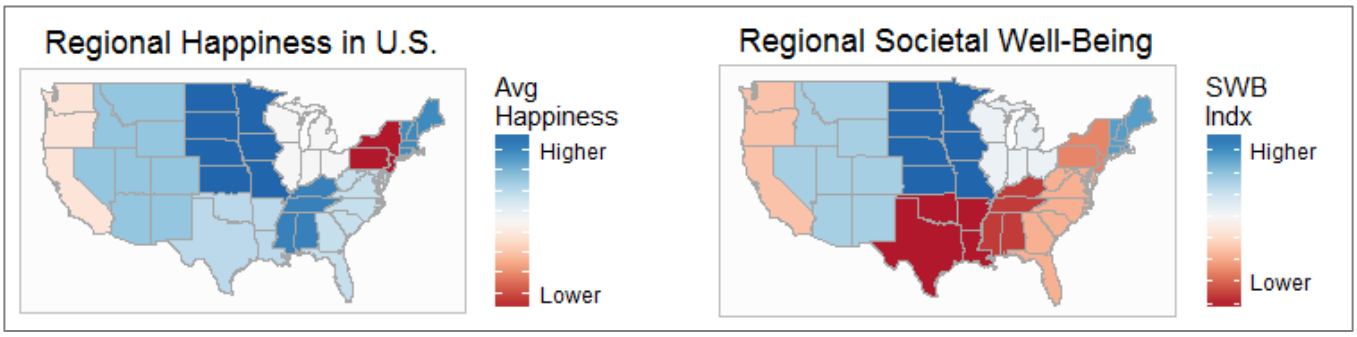

Figure 2. Comparing a simple representation of well-being (happiness scales, on left) to a more complex representation of societal well-being (on right) to derive different insights for different geographic regions in the United States.

We can also demonstrate how the model produces interesting insights in relation to political aspects of the national population, especially when considered in conjunction with temporal information. For instance, in Fig. 2 the scatterplot dots indicate national-level averages for each year of data collection (1972-2014) for each self-identified political community as measured by party affiliations (left column plots) or by ideological views (right column plots) for the simple model (top row of plots) and the complex model (bottom row). Boxes depict the middle fifty percent of the data (with mean lines) within each category, and whiskers show the range from minimum to maximum scores. The red dashed horizontal lines show overall means (across all categories). Especially interesting is how robust the results are; the general trends are qualitatively similar regardless of whether modeled with simplistic or complex representations of well-being.

C. *Monte Carlo Simulations and Predictions of Well-being*

The complex model, once derived as described in the previous section, may be used in Monte Carlo processes to explore the probability distributions associated with how potential changes in any subset of the input variables would impact societal well-being. The model can be extremely useful, for example, to government policy decision makers when the impacts of their decision alternatives could be vetted within a data-derived, model-driven trade space analysis tool. For example, Monte Carlo simulation modelers would be able to reliably quantify the effect that policy and funding decisions might have on societal well-being. Such considerations will enable next generation social analytics to generate better predictions, going beyond the prevailing social science policy of typically concluding a study upon reporting descriptive and inferential statistics.

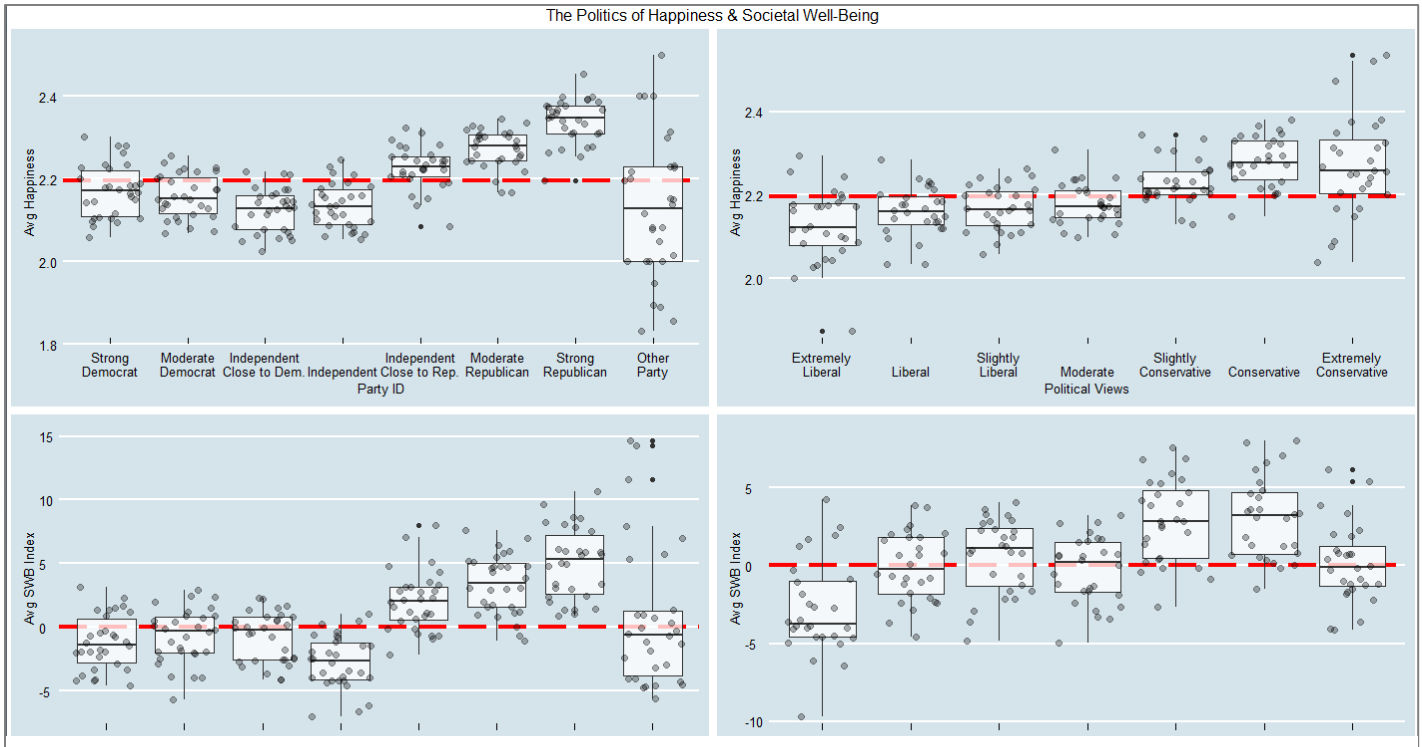

Figure 2. Aggregates of temporal data for political party and ideological views for a simplistic model of happiness versus a complex model of societal well-being

III. METHODS, TECHNIQUES, AND TOOLS FOR NEXT GENERATION SOCIAL ANALYTICS OF QUALITATIVE DATA

Next generation social scientists will also face issues related to developing methods and tools to help facilitate the collection, processing, analyzing, and visualizing of such multifaceted social data in near real-time. Our example model of individual and societal well-being is based on a static data set collected over many years. It is extremely valuable for generating structural equation models representing the interdependencies among the related input variables, and for paving the way for exploratory and predictive analyses.

Given the vast amount of qualitative data available in social media platforms such as Twitter, Facebook, and a host of blogging and microblogging technologies, it is possible to create "social sensors" which monitor important indicators of societal well-being, on massive scales, in near real-time. Traditional social science methods rely on labor and time intensive qualitative data analysis techniques to transform qualitative data into quantitative representations of affect (e.g., manually reading and coding individual text entries to determine if a person is expressing positive or negative affect). In contrast to most typical quantitative methods, qualitative data analysis methods do not easily scale up. Datasets are too large (consider the entire internet of social media, SMS/text messages, emails, blogs, etc.), and they are produced at extreme velocities (e.g., 500 million tweets per day, or status updates from 1.8 billion active Facebook users per day [10]). It is impossible for human researchers to even look at all the data, much less analysis it in a timely manner.

Whereas previous generations of Computer Assisted Qualitative Data AnalysiS (CAQDAS) software supported the traditional toolkit of qualitative researchers, i.e., sorting, searching, and annotating, the newest generation of tools is adding features powered by computerized natural language processing (NLP) and statistical machine learning (ML) techniques to enable automated rapid, massively large scale assessment of digital text, audio, video, and other multimedia traces of people's affective experiences as portrayed in their social media posts. The norm for next generation social analytics will be to employ such computational tools to facilitate blending of social media *thick data* (rich, descriptive qualitative data) with *big data* (i.e., data that is characterized by massive volume (amount of data), velocity (speed of data in or out), and variety (range of data types and sources)).

A. VADER: Automated Analysis of Affect in Social Media

VADER (**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner) [2] is a computational tool for conducting automated large scale sentiment analysis [11], [12]. Sentiment analysis is useful to a wide range of problems that are of interest to next generation social analysts, practitioners, and researchers from fields such as sociology, marketing and advertising, psychology, economics, and political science. The inherent nature of microblog content - such as those observed on Twitter and Facebook - poses serious challenges to practical applications of sentiment analysis. Some of these challenges stem from the sheer rate and volume of user generated social content, combined with the contextual sparseness resulting from shortness of the text and a tendency to use abbreviated language conventions to express sentiments. VADER is a simple rule-based algorithm and model for general sentiment analysis. In previous work [2], we compared VADER's effectiveness to eleven typical state-of-practice benchmarks for automated sentiment analysis, including LIWC [13], [14], ANEW [15], the General Inquirer [16], SentiWordNet [17], and machine learning oriented techniques relying on Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) algorithms. We used a combination of qualitative and quantitative methods to produce, and then empirically validate, a *gold-standard* sentiment lexicon that is especially attuned to affective expressions in microblog-like contexts. VADER combines these lexical features with consideration for five generalizable rules that embody grammatical and syntactical conventions that humans use when expressing or emphasizing sentiment *intensity*. We found that incorporating these heuristics improves the accuracy of the sentiment analysis engine across several domain contexts (social media text, NY Times editorials, movie reviews, and product reviews). Notably, the VADER affective sentiment lexicon performs exceptionally well in the social media domain. The correlation coefficient shows that the VADER computational engine performs as well (r = 0.881) as individual *human* raters (r = 0.888) at matching ground truth (i.e., the aggregated group mean from 20 human raters for sentiment intensity of each text-based affective expression). Surprisingly, when we further inspect the classification accuracy, we see that VADER ($F1$ = 0.96) actually even outperforms individual human raters ($F1$ = 0.84) at correctly classifying the sentiment of tweets into positive, neutral, or negative classes.

B. CASTR: Aid to Automated Topic Models of Social Text

CASTR (**C**ommon-ground **A**cquisition for **S**ocial **T**opic **R**ecognition), produces the supporting text-based information needed to establish so called *common ground*, a well-known construct from psycholinguistics whereby individuals engaged in communication share mutual facts and knowledge in order to be better understood [3], [4]. CASTR is intended to aid in *computational topic modeling* [18] by automatically acquiring this background knowledge.

Computational topic modeling techniques are used to uncover the hidden, or latent, concept-based semantic structures (i.e., topics) within text documents. Topic modeling is useful for a broad collection of activities, from automatically tagging newspaper articles with their appropriate newspaper sections (e.g., sports, finance, lifestyle, etc.) to automatically clustering like-minded social media users into groups based on the similarity of their expressed interests. Unfortunately, however, these automated approaches will sometimes infer topics that match poorly to – and are less semantically meaningful than – human inferred topics [19]. The issue is compounded when mining so-called *social text*, i.e., sparse text produced explicitly for informal social consumption (e.g., via social media, instant messages, SMS/texts, personal email, and so on where people rely on one another's common knowledge, rather than extended textual documentation, to understand intended meanings). In

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-519-7

designing and developing CASTR's algorithms, we qualitatively assess the unique characteristics of social text which present challenges to computational topic models, and which are not prevalent in other typical (non-social) text corpora like newspaper articles, scientific publications, or books. We find that a) constraints imposed by typical social media technologies, b) implicit social communication norms, and c) evolving conventions of use often confound typical computational topic modeling techniques for social text. For example, tweets are much terser than other kinds of text documents, and this sparsity is troublesome for computational topic modeling algorithms that perform posterior inference of the text. Also, tweets are often laden with a great deal of social communication "noise" (such as emoticons, emojis, hashtags, and URL links) that confuse computational models, and yet present very little trouble to humans.

CASTR leverages the concept of common ground to present a theoretically informed social and cognitive psychological framing of we refer to as the "human interpretability problem" as observed in computationally-produced topic models of text mined from social media. Additionally, CASTR employs a well-established theory from the field of Human-Centered Computing, namely Distributed Cognition (DCog) [20], [21], as a basis for mitigating the issues of developing common ground for computational topic modeling efforts. DCog is a theoretical perspective that proposes knowledge and cognition are not confined to any single individual or referent resource; instead, they are distributed across individuals, objects, artefacts, and tools in the environment, and constructed in context.

As an example of how CASTR implements the DCog inspired mitigation strategies, consider a fictitious (but representative) social media post that expresses a person's positive affective experience related to attending a musical concert at a popular venue near Atlanta, Georgia: "*Headed to Stone Mountain to see the Rolling Stones. Mick Rocks! www.rollingstones.com/band/ #StonesOnFire*". Although it is a relatively simple thing for humans to immediately understand the meaning of this social text (most Americans know who The Rolling Stones are, most people from Georgia know what Stone Mountain is, and most people understand what it means when "rock" is used as a verb in this context, even if they are not immediately sure who Mick refers to, and most people recognize the conventional use of hashtags, as well as URL links). However, the shared, socially constructed knowledge (common-ground) necessary to understand the intended meaning of the above example social text is often not readily available to computational topic models.

CASTR automatically retrieves the (previously missing) background distributed knowledge about key words, phrases, and named entities (proper nouns) within the terse text, and provides this information to the computational topic model processes. The result is a much more accurate representation of which topic(s) a particular short social media document should be belong. For example, the social text above would be appropriately grouped with music and entertainment related topics, rather than geological science related topics.

C. EAGLE-ID: Automated Demographic Profiling

EAGLE-ID (**E**thnicity, **A**ge, **G**ender, and **L**iteracy/**E**ducation **Id**entifier) automatically aids in characterizing important human social demographic features based on social media profile data. The EAGLE-ID system consists of software (currently in beta stage) which performs automatic classification of a person's ethnicity (given the person's surname), their likely age range and gender (based on their first name), and their literacy and education level based solely on information mined from the person's digital social media data (including user profile data as well as shared content). The majority of this is done via text-based computational linguistic processing (in conjunction with comparisons to data from the U.S. Census Bureau database, Social Security Administration records, and U.S. Dept. of Health and Human Services data), but it also uses computer vision for image processing on profile pictures to boost ethnicity/age/gender classification accuracy.

In addition to the obvious uses for user profiling and user modeling, the EAGLE-ID software could be useful for automatically collecting and associating demographic information with particular social media accounts, When used in conjunction with VADER and CASTR, EAGLE-ID facilitates rapid, large scale analysis of social data for use in real-time monitoring of individual and societal well-being with realistically representational complex models.

While the design and development of tools such as VADER, CASTR, and EAGLE-ID is not necessarily in the direct purview of social science, the employment and use of such tools will almost certainly be a significant part of next generation social analytics. It is already a major part of the new field of Computational Social Science. Eventually, the word "computational" will be dropped, and methods, tools, and techniques like the ones discussed in this section will be commonplace in social science research – integrated into social science education right alongside experimental study design, research ethics, and statistical analysis.

D. Crowdsourcing for Scaling-Up Qualitative Data Coding

An interesting interim step preceding fully automated artificial intelligent machine learning algorithms for conducting large scale qualitative data analyses are the emergence of digital crowdsourcing economies such as Amazon Mechanical Turk. These platforms are typically comprised of a massive, distributed, anonymous crowd of individuals willing to perform general human-intelligence micro-tasks for micro-payments, and they can be leveraged as a valuable resource for the next generation of social science research and practice. Indeed, in the past half-decade, Amazon Mechanical Turk has radically changed the way many social science scholars do research. The availability of a massive, distributed, anonymous crowd of individuals willing to perform general human-intelligence micro-tasks for micro-payments is a valuable resource for researchers and practitioners.

In other work [5], we addressed many of the challenges facing researchers using crowd-sourced platforms. Particularly, we reported on how to better ensure *high quality*

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-519-7

qualitative data annotations for tasks of varying difficulty from a transient crowd of anonymous, non-experts. Crowdsourcing has already had a significant impact on social analytics, and we believe it will continue to play a substantial role in the next generation of social analytics.

IV. CONCLUSIONS

A. A Departure from Traditional Social Analytics

The model described in the first part of this paper (c.f., Section II) differs from traditional social science in several meaningful ways:

1. *Representational complexity*: In next generation social analytics, model complexity will increase beyond what is typical for much of social science research today. Our example integrates more than 130 indicators for- and correlates of- individual and public well-being. These data are garnered from many sources, measured in numerous different units, stored using many data types at different scales representing individuals, communities, and entire societies. Just as other disciplines such as systems engineering, economics, and computer science have embraced the notion of incorporating "big data" into their typical data models, the next generation of social analytics will need to likewise expand their scope such that social analytics like the ones we illustrate are the norm, rather than the exception.
2. *Large-N and Multiple-T*: In order to achieve useful statistical power while incorporating the expanded scope resulting from increased representational complexity, and at the same time preserving broad generalization and application capacities, next generation social analysts will need to design and conduct studies with much larger sample sizes (i.e., "Large N" studies) collected over multiple instances in time (i.e., "Multiple T", or longitudinal studies). In our example, we integrate data from nearly 60,000 respondents spanning 42 years with regard to 130 different variables of interest, where each variable puts (on average) potentially 7 unique degrees of positive or negative pressure on individual or societal well-being. All told, this leverages approximately 55 million data points for our model. Such study designs will eventually become more prevalent for social analytics.
3. *Extending exploratory and predictive analytics*: Our example model lays the foundations for predictive analysis (e.g., via Monte Carlo simulations), which would be extremely useful to government policy decision makers because the impacts of their decision alternatives could be vetted within a data-derived, model-driven trade space analysis tool. For example, we would be able to answer important questions such as: *in order to improve overall community/public well-being, should government decision makers invest tax dollars in a better public transportation system, economic development program, roads, schools, or security services*? Such considerations will enable next generation social analytics to generate better predictions, going beyond the prevailing social science policy of typically concluding a study upon reporting descriptive and inferential statistics.

B. A Vision of Next Generation Social Analytics

Combining the increase in representational complexity for social science analyses described in Section II with the methods, techniques and tools described in the Section III, a vision of how next generation social analytics will be conducted begins to emerge in which large-scale, individual and national-level, near real-time analysis of the following are common:

- social media data
- mobile and GPS technology data
- personal wearable technology data
- internet of things data

In the second part of this paper, we outlined how new tools and techniques could be leveraged to marshal in the next generation of qualitative social analytics on heretofore unprecedented scales. VADER (see Section III-A) provides researchers the ability to automatically quantify both the direction (e.g., positive or negative) and magnitude of affective expressions in textual documents ranging from word-level to tome-level scales. In a matter of seconds, VADER is capable of automatically transforming millions of rich qualitative social media documents (e.g., tweets) into quantified measures of positive and negative affect for a given Twitter user. This capability alone allows us to produce a simple representation of well-being on a national scale in near-real time [2]. When we combine it with the ability to also understand the topic towards which the affective expressions apply (see the discussion of CASTR in Section III-B), we can begin to incorporate other elements of the more complex representation of well-being previously discussed.

For example, consider when a Twitter user laments (or praises) aspects of her job, her health, her family or friends, her city/community, or her financial situation. Or consider how often she might express satisfaction (or dissatisfaction) for aspects of the general political, security, or economic climate of her community or nation. Now consider how prevalent such expressions are in aggregate for all Twitter users. Next think about how many other publically available forms of such data currently exist (other social networks like Facebook and Snapchat, place-based platform Foursquare, review platform Yelp, internet chat rooms, topical blogs, and discussion forums such as Reddit). Next generation social analytics should embrace such resources, as well as the tools needed for analyzing them at internet scale.

Typically, these social media data are time-stamped, so that temporal aspects can be incorporated (c.f., [22]). Slower changing data variables such as a person's demographic characteristics (e.g., ethnicity, age, gender, literacy and education level) can also be automatically extracted from a person's social media data (see the discussion of EAGLE-ID in Section III-C). In many cases, these data can be combined with meta-information regarding the geolocated origins of the content producers, or otherwise merged with GPS, mobile, or other location-aware wearable technologies. Additional real-time assimilation of national, regional, or local unemployment rates, crime data, housing market data, inflation, consumer price index, prime rates, and gross domestic product round out

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-519-7

the capability to produce timely, realistically complex models of societal well-being like the one discussed in Section II.

C. Additional Issues and Items of Considertion

1) Model Complexity vs Model Interpretability

Increasing representational complexity in the way we discuss in Section II, while more characteristic of real-world human social behavior, is not devoid of its own issues; complex models are by their very nature more difficult to interpret. We offer a brief discussion of three avenues for mitigating the challenge of interpreting complex models. First, social science data analysts will need simple and intuitive interfaces for exploring the trade-space of the data. Such tools will increasing model transparency, and incorporating interactive data exploration will aid analysts in easily and quickly uncovering complex interrelationships within and among the variables of any complex model. Second, analysts need simple interfaces that allow them to rapidly build and assess Monte Carlo simulations regarding how potential changes in input variables impact selected response variables of interest. Third, advanced interactive data and information visualization tools will be critical for next generation social analytics to make sense of data at varying levels of aggregation and combination.

2) Ethical Considerations of Widespread Human Social Data Analytics

- Collection and continued monitoring – issues of personal privacy?
- Data ownership and use – do content producers exclusively own publicly available personal data?
- Consequences for types of algorithmic error – what are (or should be) the consequences?

3) Skill Sets and Education for NGSA

We must educate and train the next generation of social data analysts to be comfortable embracing representational complexity and incorporating methods, tools, and techniques like the ones discussed above. It will need to become standard parts of social science education, integrated into social science curricula right alongside research methods and experimental study design, research ethics, and statistical analysis.

ACKNOWLEDGMENT

The author thanks Elizabeth Williams, Dennis Folds, Molly Nadolski, and Tom McDermott for their work on the complex model used as a case study for the first part of this paper. The full published paper for that effort is yet to come.

REFERENCES

[1] C. Geertz, "Thick Description: Toward an Interpretive Theory of Culture," in *The interpretation of cultures: selected essays*, New York, NY: Basic Books, 1973, pp. 3–30.

[2] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014, pp. 216–255.

[3] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[4] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on socially shared cognition*, L. B. Resnick, J. M. Levine, and S. D. Teasley, Eds. Washington DC: APA Books, 1991.

[5] T. Mitra, C. J. Hutto, and E. Gilbert, "Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 1345–1354.

[6] E. Diener, "Assessing subjective well-being: Progress and opportunities," *Soc. Indic. Res.*, vol. 31, no. 2, pp. 103–157, Feb. 1994.

[7] E. Diener, E. M. Suh, R. E. Lucas, and H. L. Smith, "Subjective well-being: Three decades of progress.," *Psychol. Bull.*, vol. 125, no. 2, pp. 276–302, 1999.

[8] D. J. Folds and V. M. Thompson, "Engineering human capital: A system of systems modeling approach," in *Proceedings of the 8th International IEEE Conference on Systems of Systems Engineering (SoSE-13)*, 2013, pp. 285–290.

[9] T. W. Smith, P. V. Marsden, M. Hout, and J. Kim, "General Social Surveys, 1972-2014 [machine-readable data file]." NORC at the University of Chicago [producer and distributor], 2014.

[10] InternetLiveStats.com, "Internet Live Stats," *Internet Live Stats - Internet Usage and Social Media Statistics*, 2016. [Online]. Available: http://www.internetlivestats.com/. [Accessed: 09-Sep-2016].

[11] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, 2008.

[12] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA: Morgan & Claypool, 2012.

[13] J. W. Pennebaker, M. Francis, and R. Booth, *Linguistic Inquiry and Word Count: LIWC 2001*. Mahwah, NJ: Erlbaum Publishers, 2001.

[14] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC.net, 2007.

[15] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," NIMH Center for the Study of Emotion and Attention, Center for Research in Psychophysiology, University of Florida, Technical Report C-1, 1999.

[16] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie, *General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press, 1966.

[17] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in *Proc. of LREC*, 2010.

[18] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[19] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models," in *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 2009.

[20] J. Hollan, E. Hutchins, and D. Kirsh, "Distributed Cognition: Toward a new foundation for human computer interaction research," *ACM Trans. Comput.-Hum. Interact. TOCHI*, vol. 7, no. 2, pp. 174–196, 2000.

[21] E. Hutchins, "Distributed Cognition," in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 2068–2072.

[22] C. J. Hutto, S. Yardi, and E. Gilbert, "A Longitudinal Study of Follow Predictors on Twitter," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France, 2013, pp. 821–830.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-519-7