Integrating Machine Learning (ML) and Grounded Theory Research: *Exploring Information Privacy and Doctor-Patient Relationship*

Emergent Research Forum (ERF)

A. F. Salam
ISSCM Dept., Bryan School of Business and Economics, UNCG
amsalam@uncg.edu

Shabnam Nahar
Dept. of Public Health and Informatics, Jahangirnagar University, Dhaka, Bangladesh
sathy303@gmail.com

Sabbir Pervez
Eye Care Project, Manabik Sahajjo Sangstha, Dhaka, Bangladesh
sabbirpervej@gmail.com

Abstract

We attempt to develop a research method that incorporates Machine Learning and Grounded Theory (GT) Research to analyze massive text data. Traditional GT is limited by human cognitive limitations wherein processing massive text data is nearly impossible to interpret and analyze. On the other hand, Machine Learning techniques are suitable for massive text analysis but lacks human judgement, intuition and interpretive capacities. By combining GT with Machine Learning, we attempt to extend traditional GT with Machine Learning so that researchers are able to analyze massive text to develop novel and interesting theories. However, ontologically and epistemologically GT and Machine Learning have different origins. We propose to incorporate Machine Learning with GT using 'Emergence" as central to Grounded Theory research. We argue that Machine Learning Model serves in place of human computational nodes where teams of researchers are used in large scale GT research projects.

Keywords

Machine Learning, Grounded Theory, Large Text Analytics, Doctor-Patient Relationship, Information Privacy

Introduction

How do we integrate Machine Learning and Text Analytics (MLTA) with rich tradition of interpretive research such as Grounded Theory (GT) while maintaining the epistemological and ontological foundations to generate meaningful theory in the context of large text data (often in the millions of lines of text)? How this approach can be used to extend conceptualization of information privacy in Doctor-Patient relationship and eHealth incorporating the notion of 'Body' from cultural anthropology? Increasingly, information systems (IS) scholars investigate research questions related to social media phenomenon that require analyzing enormous amounts of data (Berente et. al, 2018). Available massive datasets such as product and service reviews (one example dataset contains approximately 24 million reviews), newspaper archives, twitter, Facebook data etc. offer enormous opportunities for investigating interesting research questions and development of novel theories. Although a significant number of information systems (IS) studies have used Machine Learning (ML) in auctions, social media research etc., this set of research do not address how to integrate traditional interpretive research approaches with Machine Learning Algorithmic techniques in analyzing massive text datasets. Interpretive research requires close examination of data and careful interpretation of the context to extract meaning from text. IS scholars can develop rich, thick descriptions (Geertz, 1973) of interesting phenomenon using interpretive method such as Grounded Theory. This careful and close examination of data (text) is time-consuming and thus limits how much data can be practically analyzed and thus posing a significant challenge when faced with massive text data sets available in social media. Some of these datasets can be in the range of millions of lines of text. Too much data is nearly impossible to be processed by humans even if a reasonable number of teams of trained scholars are used. In this context, it is critical that IS scholars begin to consider how machine learning (ML) solutions specifically those that relate to natural language processing (NLP) and text analytics might augment traditional interpretive research methods such as Grounded Theory.

One might argue that ontologically and epistemologically Grounded Theory and Machine Learning Text Analytics (MLTA) are likely to be divergent as these approaches to analyzing text have different origins and serve different research objectives. "Moreover, an appeal to computational objectivity may embody fundamentally different epistemic commitments than those at work in interpretivist approaches such as Grounded Theory (Baumer et. al. 2017, pp.1). However, recent research (Berente et. al, 2018) point to promising integration among these diverse research methods. Additionally, research in sociology (Nelson, 2017; Nelson et. al, 2017) and information sciences (Baumer, et. al, 2017) have made limited attempt at integration between Grounded Theory (GT) Research Method and Machine Learning Text Analytics (MLTA) and comparison among the methods. This set of research have made significant contribution and have laid the rudimentary foundation for further investigation to integrate Grounded Theory and MLTA.

Although recent studies (Berente et. al, 2018; Baumer et. al, 2017; Nelson, 2017; Nelson et. al. 2017) make important contribution to advance our knowledge, they do not address the ontological and epistemological concerns underlying the integration of Grounded Theory and Machine Learning. Without a clear understanding of the underlying ontological and epistemological concerns, scholars attempting to answer interesting research questions and subsequent theory building based on integrating GT and ML to analyze massive datasets are likely to analyze data, draw conclusions or build theories that may not be appropriate. Absence of considering and addressing ontological and epistemological foundation of each method is likely to lead to confusion about each method employed and subsequent misinterpretation of research results and consequent questionable theory building. A state undesirable to the IS and other scholar communities. In this research, we address this gap in IS literature. We strongly believe that the IS discipline is uniquely situated and should be at the forefront of advancing both understanding and knowledge in how to integrate interpretive methods (such as GT) and Machine Learning in analyzing massive text datasets for novel theory generation.

Human interpretation is critical yet due to bounded and cognitive limitations, we cannot process massive text data sets. On the other hand, machine learning algorithms can process massive text data sets, yet these systems lack human intuition, judgement, interpretive capacities, etc. Scholars (Berente et. al, 2018; Baumer et. al, 2017; Nelson et. al, 2017; Nelson, 2017) believe and have laid the rudimentary foundation that it is possible and promising to incorporate Grounded Theory Research Method with Machine Learning Text Analytics (MLTA). In this research, we build on that nascent yet important foundation to help scholars build the next generation of new IS theories that are so critically needed. In this research, we use the fundamental concept of "emergence" in Grounded Theory and its philosophical foundations to address our research questions. Emergence is central to GT as it alludes to the nature of "theory building" in GT as "theory" emerging from data with the skills, experience and interpretive capability of the analyst or teams of analyst.

As part of future research, we are applying the proposed framework of integrating Machine Learning with Grounded Theory in the context of Doctor-Patient relationship and information privacy and eHealth. In this research, we delve into Cultural Anthropology and the notion of "Body" especially in the context of female patient and Doctor-Patient relationship. The body is the material basis of human beings which defines our existence. On the one hand, the body is a part of human existence, which the individual cannot choose freely. Between the given body on the one hand and intentional body management on the other, body culture develops in a process, which is both historical and collective (Eichberg, 2007). According to Eichberg (2007), what is neglected is the body as a field of dynamic human interaction and that of movement. Scheper-Hughes and Lock (1987) have defined 'three bodies' at three levels of sociocultural analysis. These are: the 'individual body', 'the social body' and the 'body politic'. While all three levels of conceptualizing the body are important and useful, this research focuses, in particular, on the 'lived experience' of the individual body especially female patient and Doctor, and the ways that medical knowledge and practices create discourses in this context of Doctor-Patient relationship. Additionally, this study investigates implications for eHealth management given the privacy and sensitivity and identity implications of the notion of 'body' in the medical context. We investigate the notion of 'body' from a female patient perspective on interview data collected from patients, doctors and nurses.

Proposed Framework Integrating Grounded Theory, Constructionist Emergence and Machine Learning Text Analytics

In this research, we adopt the constructionist emergence with a critical realist ontology and a relativist epistemology. In this approach, emergence depends on the interaction between the researcher and data. The data influence how the researcher constructs the emerging theory and the emerging theory influences how the researcher interprets data as well as subsequent collection, analyses and interpretation and re-interpretation of data (Charmaz, 2006). Due to the adoption of a critical realist ontology the viewed exists outside of the mind of the viewer and the meaning is constructed through an interaction between the viewer and the viewed. Having a critical realist ontology also implies that there is a real world which the participants and researcher are able to access in bits and pieces as embedded and captured in data. The bits and pieces come together to form and explain the data as the researcher uncovers patterns that inform the emergent theory. We incorporate Machine Learning Text Analytics (MLTA) as an extension of the fundamental tenets of Grounded Theory Research Method (Charmaz, 2014; Corbin and Struass (2008); Orlikowski, 1993, 1996; O'Reilly et al., 2012) as outlined below that involve the following non-linear steps conducted throughout data collection, analysis and writing: (1) the constant comparative method, (2) theoretical coding, (3) theoretical sampling, (4) theoretical saturation, (5) theoretical sensitivity.

At this stage, we conceptualize the role of MLTA at the first step – the constant comparative method. This stage involves the simultaneous coding and analysis of data (Glaser and Strauss, 1967). Within the constant comparison stage, all new data are compared with earlier data iteratively to enable adjustment of theoretical categories based on the ongoing analysis and the goal is "comparing incident to incident and then incident to concept for the purpose of generating categories and saturating their properties" (Glaser, 2001, pp. 185). At the constant comparative stage, the researcher gathers data pertaining to the phenomena of interest as per the research questions. As data is collected, initial analysis is conducted to classify relevant data into categories as pattern embedded in data becomes more transparent to the researcher. This constant comparison continues and the researcher observes for confirmation of already coded categories and also for new possible emerging categories. This cycle of category construction moves in cycles – verifying and reinterpreting and analyzing new data as it becomes available. This manual approach of constant comparison is limited by human bounded rationality and information overload. If the researcher faces a large data set, the researcher may employ a team of coders to code the initial data set. Each coder under this approach will use the researcher defined codes that shows the categories and exemplar sentences or text to help them code new data into specific categories already defined by the researcher. They will likely confirm the categories in analyzing new data or might identify potential for new categories in the data. This allows the researcher to offload the cognitive burden to other research assistants (coders). As categories are confirmed or potential new categories are identified, the researcher reaches a position to begin to observe the emergence of rudimentary theory based on the patterns embedded in the text data. In this context, one can conceptualize each coder (research assistants) as human computational decision nodes. By human computational decision nodes, we mean that human computational or intelligence capability of coders (research assistants or other researchers) is applied to code new data using the already established codes and text exemplars supplied by the researcher. In this context, the coders are analyzing data and categorizing those data essentially extracting the embedded pattern in the text data.

In this research, we conceptualize MLTA as computational nodes similar to human computational decision nodes where the researcher provides the categories and corresponding exemplar text data to train the machine learning model. A part of the researcher supplied exemplar data is kept aside in order to evaluate the performance of the trained machine learning model. Once the researcher is satisfied with the performance of the trained machine learning model (based on machine learning model performance parameters), new data is supplied to the trained model to search for and confirm already established categories in the new data similar to what the human coders as human computational decision nodes would accomplish. In this context, MLTA assumes the role of team of human coders. We do recognize that MLTA as pure computational nodes do not have the intuition, interpretive capacities of human coders but MLTA also does not have the human frailty of bias, errors in judgement or misinterpretation or misreading of the codes and exemplar text data. More of these characteristics will be explored in the full paper are beyond the scope of this extended abstract. In the following section, we outline our Proposed Framework in integrating Grounded Theory with Machine Learning Text Analytics (MLTA). In the proposed framework, the researcher begins with human interpretation of the available text data and manually constructs the initial codes and categories and exemplar text data corresponding to each category. In the next step, machine learning training data file (TD1) is created which holds the categories and corresponding text exemplars for each category. These text exemplars can be specific sentences or groups of sentences depending upon how the categories are coded or defined by the researcher. In our framework, we choose Supervised Machine Learning Text Analytics. Supervised machine learning works with researcher supplied text and categories. It does not explore on its own any categories unlike Unsupervised Machine Learning used in earlier research by Baumer et al. (2017), Nelson et al. (2017) and Nelson (2017). The primary deficiency of unsupervised machine learning from our perspective is that unsupervised machine learning cannot use researcher defined categories which clearly violates the fundamental tenets of Grounded Theory where the researcher analyzes data, interprets and constructs categories based on the research question of interest.

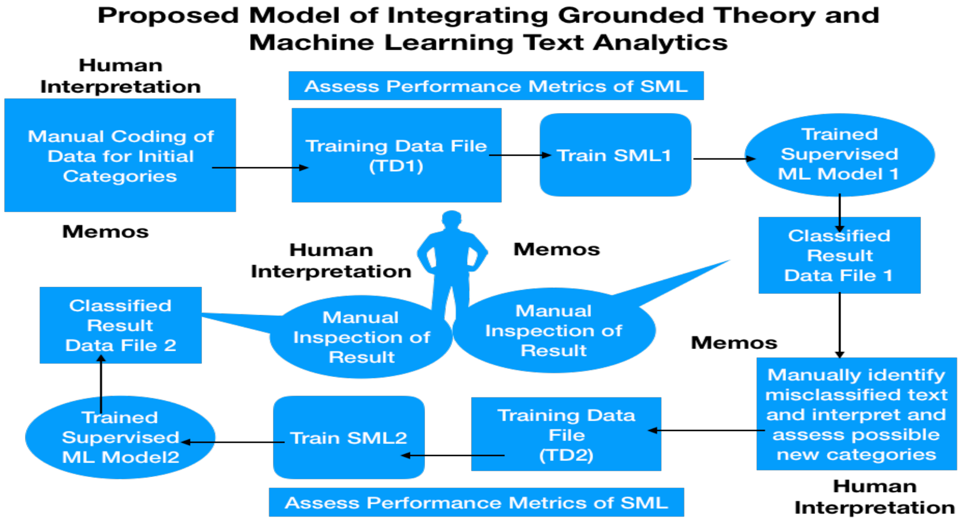
Proposed Model of Integrating Grounded Theory and Machine Learning Text Analytics
Human Interpretation
Assess Performance Metrics of SML
Manual Coding of Data for Initial Categories
Training Data File (TD1)
Train SML1
Trained Supervised ML Model 1
Memos
Human Interpretation
Memos
Classified Result Data File 1
Classified Result Data File 2
Manual Inspection of Result
Manual Inspection of Result
Memos
Manually identify misclassified text and interpret and assess possible new categories
Trained Supervised ML Model2
Train SML2
Training Data File (TD2)
Assess Performance Metrics of SML
Human Interpretation

Figure 1 Framework Integrating Machine Learning and Grounded Theory Research

These fundamental requirements cannot be met with unsupervised machine learning. Interestingly, supervised machine learning does not have these limitations and is therefore used in our proposed framework. Once the training data file (TD1) is available it can then be used for training the machine learning model. Trained machine learning model is then evaluated with a new set of similar data but previously unseen by the trained model. Once the performance metrics for the trained model (such as precision and recall metrics and confusion metrics) are all satisfactory, the researcher can then feed new data for categorization by the trained model. The output from the trained model then can be analyzed by the researcher for confirmation of earlier codes and also for identifying potential new categories. This is very similar to the human coders as discussed earlier. To continue with this interative process, the researcher can create a new training data file (TD2) including new categories and exemplar text. This new TD2 file can be used for training a new machine learning model on a new data set to confirm and discover potentially new categories. This cycle can continue until saturation is reached and no new categories emerge from the text data. Interestingly, due to the incorporation of MLTA within the Grounded Theory method, the researcher is no longer limited to analyzing only a handful of text data due to human cognitive limitations. Now the researcher is able to analyze millions of text data using MLTA without sacrificing the rigor and interpretive strength of Grounded Theory. In this overall process, memos can be written by the researcher for new findings or new categories and most importantly about the various elements of the emerging theory. In this research, we use the constructionist emergence paradigm with critical realist ontology and a relativist epistemology. Critical realist ontology allows the researcher to maintain a distance from the data identifying that reality exists independent of the researcher so that the "bits and pieces" of observed and collected data captured or embeds the phenomena of interest. Additionally, the relativist epistemology allows the researcher to construct the emergent theory as a participant in the theory construction process where the researcher influences the data and the data influences the researcher. This interaction leads to emerging theory that is intimately tied to the data and the interpretive capacity, judgement, experience and intuition of the researcher. MLTA remains faithful to this constructivist emergence paradigm by providing the analytical faithfulness by following the researcher supplied code and categories and unearthing the researcher specified pattern embedded in the text data.

Conclusion and Future Research

We are applying the Proposed Machine Learning and Grounded Theory approach to investigate the notion of 'Body' in the context of Doctor-Patient relationship using cultural anthropology. The notion of 'Body' signifies a deep level of 'sensitivity' and personal identity that is often missing in the extant literature on information systems privacy especially in the context of eHealth.

Exploring Doctor-Patient Relationship: Extending the Conceptualization of Information Privacy with 'Sensitivity' Emanating from the Notion of 'Body' in Cultural Anthropology

To develop this notion of 'sensitivity' in the context of 'Body', we have collected extensive interview and archival data in the context of Doctor-Patient relationship. This huge quantity of data cannot be analyzed effectively using even teams of researchers. To apply Grounded Theory approach to the large collection of text data, we are employing the Proposed Framework (See figure 1) and the Machine Learning and Text Analytics software called 'NoCodeMachineLearningApp' developed by the first author which requires no coding in Python or R. Following the Proposed Framework (Figure 1), the authors have developed Grounded Theory coding scheme reflecting the initial codes representing the notion of 'Body'. Initial set of codes have been used with the 'NoCodeMachineLearningApp' to analyze initial set of collected data. The authors are in the process of analyzing the output of the proposed framework and to develop emergent theory.

REFERENCES (Abridged)

Blei, David M.2012. "Probabilistic Topic Models." Communications of the ACM 55: 77-84.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 1, 993–1022.

Carley, Kathleen. 1994. "Extracting Culture through Textual Analysis." Poetics 22: 291-312.

Charmaz, K. (2006). Constructing grounded theory: A practical guide through qualitative analysis. London: Sage.