The Grounded Text Mining (GTxM) Framework

Overview and Framework Specification

By Babatunde Kazeem Oladejo

July 2024

Overview

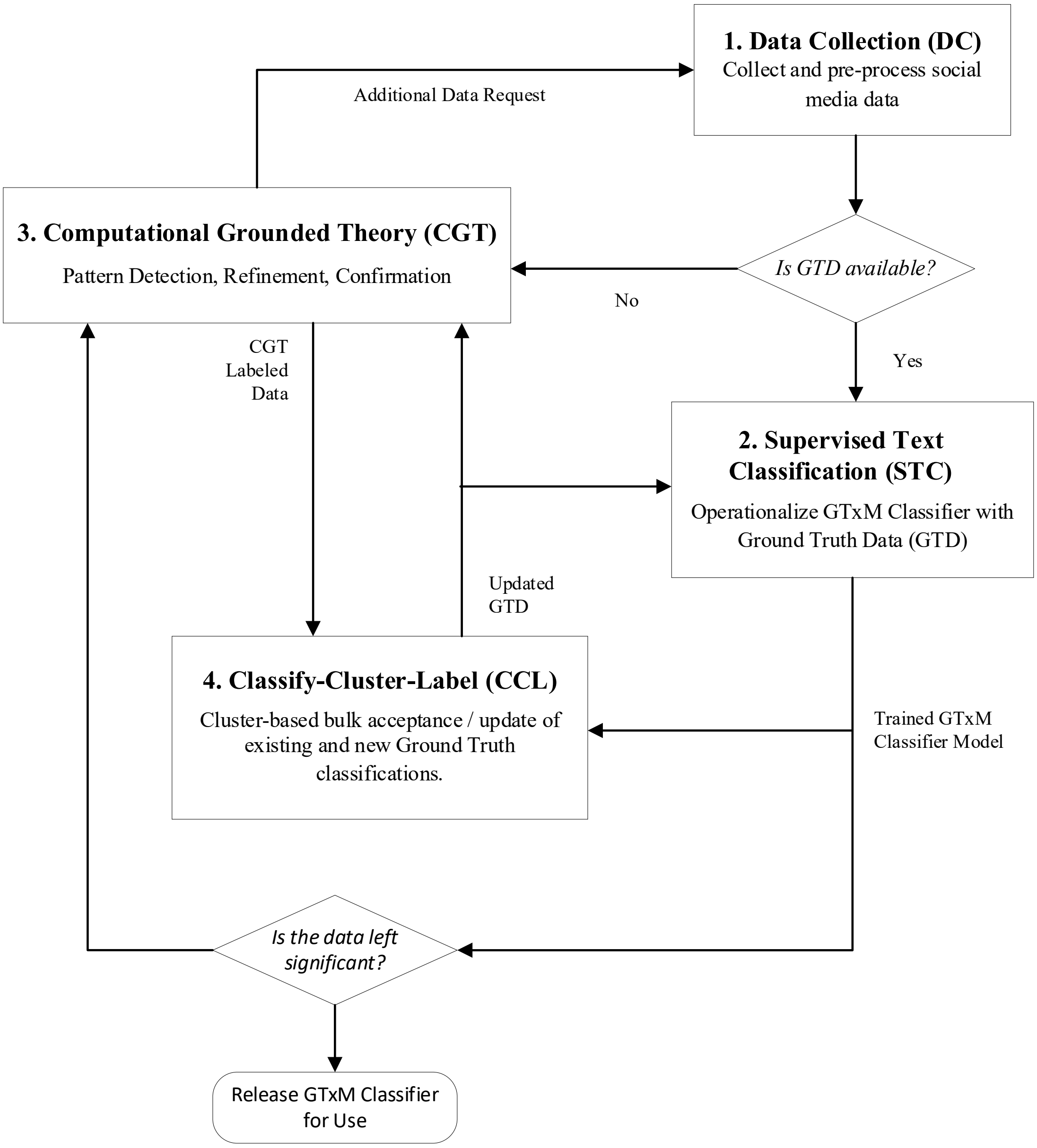
The thesis proposed Grounded Texting Mining (GTxM) framework provides a quality controlled, theoretically prudent approach to the selection and classification of social media records from a collection of social media content. The GTxM framework consists of text mining components that enable the derivation of categorical patterns and topical classifications from unstructured text [1], whilst applying the thorough and systematic qualitative approach of grounded theory [2]. The Grounded Text Mining framework consists of the following four sub-frameworks (see Figure 1 for flowchart):

A. **Data Collection (DC):** this sub-framework uses automation methods to extract unstructured social media data from news media sources. Additionally, while in DC, the researcher explores the data for possible presence of Ground Truth Data (GTD). If GTD is available, the process flow goes to STC, else CGT.

B. **Supervised Text Classification (STC)**: is a sub-framework that creates and maintains a machine learning based classifier capable of the prediction and operationalization of social media record classifications, called the GTxM Classifier. The GTxM classifier uses an ensemble of 3 supervised ML algorithms to predict labels for social media data based on the training with GTD. The STC process releases a version of the GTxM classifier for operational use or with the CCL sub-framework's classification task.

C. **Computational Grounded Theory (CGT):** is a 3-step grounded theory compliant sub-framework that performs inductive discovery of new and emergent classifications from big data through pattern detection, refinement, and confirmation. The process releases CGT labeled data to the CCL sub-framework.

D. **Classify-Cluster-Label (CCL):** is a sub-framework that uses an unsupervised clustering method and the GTxM Classifier to enable bulk semi-automated, semi-manual validation of CGT generated labels. The CCL sub-framework is also used to update existing GTD labels (if required). The CCL process releases labeled GTD to the STC sub-framework for incrementally training of the GTxM Classifier or to the CGT sub-framework for the refinement of CGT generated labels.



1. Data Collection (DC)
Collect and pre-process social media data
Additional Data Request
3. Computational Grounded Theory (CGT)
Pattern Detection, Refinement, Confirmation
Is GTD available?
No
Yes
CGT Labeled Data
2. Supervised Text Classification (STC)
Operationalize GTxM Classifier with Ground Truth Data (GTD)
Updated GTD
4. Classify-Cluster-Label (CCL)
Cluster-based bulk acceptance / update of existing and new Ground Truth classifications.
Trained GTxM Classifier Model
Is the data left significant?
Release GTxM Classifier for Use

Figure 1: Grounded Text Mining (GTxM) Process Flowchart

The GTxM Continuum

Philosopher and mathematician Bernard Bolzano defined the continuum as composed of simple, non-continual objects (points in time, space or substances) which have at least one neighboring inter-related object at every distance, no matter how small the distance may be [3]. The GTxM continuum is structured in this manner, in a simple format that allows the four sub-frameworks to be continuously engaged for the discovery and generation of (record categories) theories in a grounded text mining paradigm.

In the GTxM continuum, the framework components work in cooperation with each other, without an explicit start point and any step could be skipped for a specific cycle or pass. For example, if the researcher already possessed ground-truth data (in the case of using an existing dataset), DC can be skipped, and the process can be started at STC. And while in the GTxM continuum, the researcher has the flexibility of using minimum or no output from a sub-framework to access another sub-framework, and loop back as many times as necessary to meet the research design objectives.

The GTxM continuum experiments are organized as passes in this thesis to ensure orderly reporting. Since a GTxM continuum pass can start at practically any step, the important delimiter for passes is the end of the pass. A GTxM continuum pass ends with an activity that generates new or updates the GTD and results in the incremental training of the GTxM Classifier. The release of a new version of the GTxM Classifier prompts the flowchart decision question: 'is there significant data left?' if yes, a new continuum pass is invoked, otherwise, the GTxM process flow ends.

References

[1] U. Y. Nahm and R. J. Mooney, “Text Mining with Information Extraction,” p. 8.

[2] K. Charmaz and R. Thornberg, “The pursuit of quality in grounded theory,” *Qualitative Research in Psychology*, vol. 18, no. 3, pp. 305–327, 2021, doi: 10.1080/14780887.2020.1780357.

[3] V. Petrov, “The Conception of Bernard Bolzano About the Continuum and the Achievements of Contemporary Mathematics,” *Arhe*, no. 2, pp. 101–111, 2004.