Journal Title
XX(X):1–10
©The Author(s)
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Computational grounded theory revisited: from computer lead to computer assisted text analysis

Hjamar Bang Carlsen and Snorre Ralund, R&R at *Big Data and Society* - Please do not share without permission

Abstract

The size, variation in both meaning-making and populations that characterize many contemporary text data demands research processes that support both discover, interpretation and measurement. We assess one dominate strategy within the social science which takes a computer lead approach to text analysis. The approach is coined Computational Grounded Theory. This strategy we argue rely on a set of unwarranted assumption: that unsupervised models returns natural clusters of meaning, that the researcher can understand signs with limited immersion and that indirect validation is sufficient for ensuring unbiased and precise measurement. In response to these critiques we develop a framework which is computer assisted. We argue that our reformulation of computational grounded theory better aligns with principles within both Grounded Theory, anthropological theory generation and ethnography.

Keywords

Computational text analysis, interpretation, mixed methods, grounded theory, digital social research

Introduction

Textual and pictorial communication is with the digitization of many practices becoming an ever more prominent part of social interaction and communication in contemporary societies. With the rise of social media we now have communication on culture, politics and social issues from a large part of the population instead of merely elite actors. While the rise in population and practices that produce text is an important new data resource for the social sciences, it also poses additional challenges. We know that different populations speak and write differently, we know that text takes on different meanings in different contexts. In many data sets we have little prior knowledge of which populations and practices that have produced the texts. Furthermore, the size of the text corpus makes it impossible to read manually. This puts large demands on text analysis methodology that needs to ensure *discovery* of categories within the data, the *grounding* of the interpretations, correct *classification* of content, efficient and valid scaling of the classification, and it needs to ensure *good and unbiased* measurement of categories. In this article we seek to contribute to the development of such a methodology by stressing an computer assisted as opposed to what we argue is a computer lead approach to text analysis.

Two approaches to computational text analysis is commonly juxtaposed in an attempt to overcome the above identified tasks: a supervised approach and an unsupervised approach (Grimmer and Stewart (2013); Evans and Aceves (2016); Nelson et al. (2018); DiMaggio (2015)). In the supervised approach, common within computer science, machine learning models learn from humanly annotated text. Here the human coder is the *ground truth* that automated classification is trained on and evaluated against. Supervised approaches in other words rely on manual coding. The practices of coding and quantifying text has been critiqued for a number of pitfalls within the social sciences: 1) coding assumes that categories are known while they are mostly not 2) that human coder are biased and unreliable(Nelson (2020); Lee and Martin (2015); DiMaggio et al. (n.d.)), 3) and that coding hides away the most vital part of working with text material namely accounting for and justifying the finale interpretation of documents(Biernacki (2012)). Partly for these reasons, unsupervised machine learning methods have been much more popular in the social sciences. Unsupervised methods for text classification, especially topic models, have been praised for aligning with dominant assumptions within cultural sociology and qualitative inquiry: Inductively deriving categories from the data, allowing for the polysemy of words and the heteroglossia of documents, assuming the relationality of meaning(DiMaggio et al. (n.d.); Nelson (2020)). This has also lead authors to claim that there is a strong relation between grounded theory(GT) and unsupervised procedures(Nelson (2020); Baumer et al. (2017)). This relation between unsupervised procedures and GT is a potentially fruitful starting place for thinking about machine anthropology. This because GT shares many concerns with different aspects of anthropology, developing theory from systematic qualitative inquiry(as opposed to only testing theories), dissatisfaction with overly abstract and generalizing theories being forced upon data, and a commitment to understanding and taking local

[1] [2]

Corresponding author:

Email:

knowledge's and categories into account then developing ones analysis.

Nelson (2020) most clearly articulates the methodological framework behind the use of unsupervised methods and qualitative inquiry coining it *computational grounded theory*. She details 3 steps pattern discovery, pattern refinement and pattern confirmation. Key to computational grounded theory is that patterns are located by the computational model and not the human coder, it is *computationally lead* to ensure against the biased and constrained researcher. The researcher can then judge the usefulness of the categories returned by the model given substantive interests and read paradigmatic text to gain an in-depth understanding of the category and refine the interpretation. Lastly, Nelsons argues, in line with many other, that pattern confirmation can be done by correlating the models measure of the category with some other variables indicating the same category. In its present form computational grounded theory relies on the assumption that 1) unsupervised models return natural clusters of meaning 2) that the researcher can learn from minimal immersion 3) that the best way of validating a classification is through indirect measurement.

In this article we start by assessing the plausibility of each of these assumptions and reformulate the framework of computational grounded theory on the basis of this assessment. Furthermore, the article asks if the computational grounded theory still has a close relationship with anthropology in its reformulation. We argue that unsupervised methods such as topic models can not be taken to be local natural clusters of meaning. Our first argument has to do with LDA topic models specifically and their instability and tendency to lump together things that do not belong together. Our second argument is more principled, we argue that we can not assume to know how word patterns map on to meaning and therefore can't trust unsupervised models to locate patterns of meanings within text. This has the consequence that we can't let discovery be computer lead. Furthermore, the idea that the human researcher can read only the paradigmatic text retrieved by the model in order to provide an holistic and interpretatively valid account goes against much qualitative research that argues that there is a substantial phase of learning where the researcher comes to terms with how signs are used and what they refer to theoretically within a specific field site or regarding a certain type of action. This means that the researchers ability to read and understand correctly can't be assumed, but must be ensured and accounted for. Lastly, we present evidence that we can't rely on indirect validation because it does not ensure precise nor bias free measurement. Our critique of computational grounded theory leads us to an reformulation which instead of being computer lead is computer assisted. We develop the central steps in a research framework we call computer assisted learning and measurement(CALM) which goes from discovery, grounding, classification, validations and measurement.

The article starts with clarifying the relation between GT and strands of anthropology. The following section runs through and assess the different aspects of computational grounded theory, moving from discovery to measurement. In the next section we will introduce our own framework and develop what we call computer assisted text analysis, again moving through the various steps of the analysis. The paper end with a discussion and a conclusion.

Grounded theory and anthropology

In this section we'll introduce to some of the central ideas behind Grounded theory and clarify it's similarities and differences to parts of anthropology. Especially, we'll argue that grounded theory and anthropology share 1) a commitment to theory generation from systematic qualitative inquiry, 2) a critique of overly general and detached theories based on logical deduction and 3) finally a commitment to understanding the point of view of those that they study.

Grounded theory(GT) in the formulation of Glaser and Strauss (2009) is an approach to generating theory that focuses on emergent theory generated from data analysis, as opposed to theory generated from logical deduction from prior assumptions. The latter according to Glaser and Strauss leads to too many opportunistic uses of theory that do not fit the data (Glaser and Strauss 2009, 6). As an alternative to theoretical deduction, Glaser and Strauss argue that theory should be ”systematically worked out in relation to the data during the course of research. Generating a theory involves a process of research.”(Glaser and Strauss 2009, 6). A central part of GT is theoretical sampling where cases are chosen in respect to developing the category or hypotheses of interest. The goal is not empirical generalization, but rather theoretical saturation(see also Small (2009)). Hence, theoretical sampling concerns what cases to compare for a specific theoretical propose(Glaser and Strauss 2009, 47). Three points of comparison are vital to control according to Glaser and Strauss: populations, conceptual level and degree of similarity. Importantly, not because one has to compare things that are alike, but rather because theory development is dependent upon how one compares across these three parameters. The method used is the constant comparative method which analysis each instance of category with prior instances in order to get at its defining features, central properties and relations to others categories. Typically, the constant comparative methods relies on open coding of instances working towards a integrated theory. In this manner GT combines the systematic coding procedure of more traditional quantitative content analysis, typically aimed at theory testing, with aim of theory driven qualitative work aimed at developing theory, the latter typically done without any systematic data analysis(Glaser (n.d.)).

The strongest relation between GT and anthropology regards theory generation. Both GT and many strands of anthropology sees overly general and abstract theories as highly problematic. While anthropologist, more than GT, has done so for ethical and political reasons, both stress the analytical problems with ill fitted abstract theories. This concern is a central motivator for anthropology's strong tradition for developing theories from their field work, as opposed to merely testing theories. Radicalized in the ontological turn (Holbraad and Pedersen (2017)), anthropology takes the ethnographic field site as the source of their analytical categories rather merely their object. In this process no prior ontological assumptions are sacred(expect maybe for prioritizing the immediate rather than the mediated, as argued by Kockelman (2017)). Rather, precisely generating new, often radical, conceptualizations in dialog with ones material is often seen as the goal. From this perspective it is clear that GT and parts of anthropology are similar in their ambition to generate theory from qualitative inquiry.

Although seldom stated explicit, symbolic interactionism influence on GT and its dictum ”people act towards things on the basis of the meaning they attribute to them” makes GT committed to understanding meaning-making within their data in order to develop theories that fit the social situation studied. This is, of course, one of the essential tasks of fieldwork, namely to understand the ”natives point of view”(Malinowski (2002)). In some formulations, this is the main task of anthropology. What has been referred to as cultural translations. By most it is seen as requirement for an ethnography that the ethnographer understands and to some extent lives up to the requirements to interaction in their field site, before any analytical statements are made(Lichterman (2017)).

Below we will compare computational to non-computational grounded theory especially regarding the phase of discovery and interpretation of categories. However, computational grounded theory and the wider use of topic models in the social science also uses topic models as a measurement device. The output of such models are used as either the independent or dependent variable in regression analysis (Evans and Aceves (2016)). Therefore, we also have to evaluate computational grounded theory as a strategy for producing precise and unbiased measurement. The ambitions of Computational Grounded Theory to both interpret and measure makes perfect sense given the large textual data with interesting metadata(time, organization aso.) that are becoming increasingly available for social science research. These data contribute to social science with an ability to support both in-depth interpretation of selected documents and large scale analysis of socio-cultural patterns (Author).

The logic of computational grounded theory

In this section we run through the 3 different steps of computationally grounded theory: discovery of patterns, interpretations and grounding of the patterns and pattern validation.

Discovery

When confronted with large text data researchers have the challenge of discovering what categories are in the data and interpreting large amounts of text. A central claim in CGT is that this dual challenge can be solved by making discovery computationally lead. According to Nelson, the role of computational model in *pattern discovery* is to reduce the messy, complicated text into a simpler, more interpretable list or network of words (Nelson 2020, 7). This allows for relevant categories to emerge from the data that the researcher, due to their own preconceptions or the complexity of the text, had not considered. This prevents the researchers from introducing their own biases into the material. The simpler representations are then interpreted and categorized in a fashion similar to normal content analysis, but, Nelson argues, this process is ”fully and immediately reproducible” (Nelson 2020, 13), unlike human-coded text. In sum, computer-assisted *pattern discovery* that uses actual frequencies of the (co-)occurrence of words to locate ideas or concepts in the data constitutes, according to Nelson, a more ”reproducible and scientifically valid grounded theory” (Nelson 2020, 13). The task of constant comparison from grounded theory aimed at discovering new categories from the data is in CGT done by an algorithm that takes word distribution across documents as input in order to detect patterns. Two important assumptions are made when using topic models, such as latent Dirichlet allocation (LDA) topic models, for discovery. The first assumption is that the model returns actual clusters of co-occurring words. The second assumption is that patterns in co-occurring words map on to meaning. We will discuss each assumption in turn.

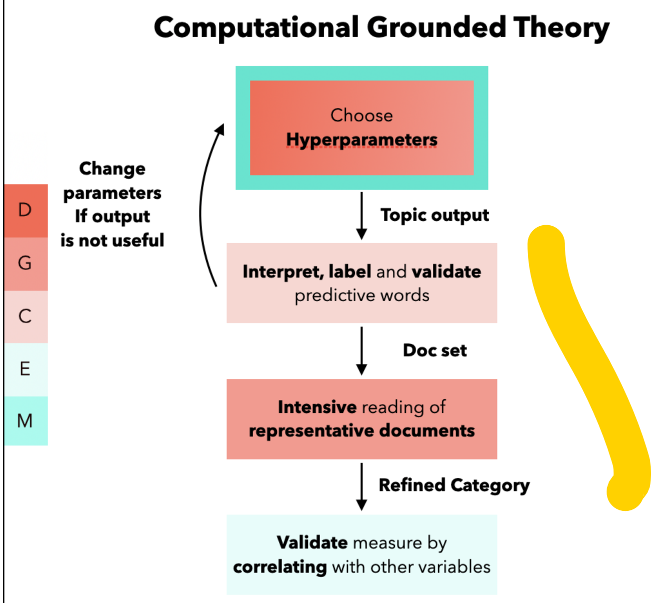
Computational Grounded Theory
Choose Hyperparameters
Topic output
Interpret, label and validate predictive words
Doc set
Intensive reading of representative documents
Refined Category
Validate measure by correlating with other variables
Change parameters If output is not useful
D
G
C
E
M

Figure 1. This is a illustration of the workflow of computational grounded theory. The different colors indicate what more general task is being handled at the different stages of the research process. D stands for Discovery; G for grounding; C for Classification; E for evaluation and validation of measurement.

A typical imaginary used to justify the use of topic models is that they cluster words together based on their co-occurrence in text DiMaggio et al. (n.d.). The co-occurrence thesis is common in quantitative text models(in linguistics its typically referred to as the distributional hypothesis). Often the linguist Robert Firth is quoted for saying that ”words shall be known for the company they keep”. This is taken to mean that words should be understood through the words that they co-occur with and not in isolation. The place of this co-occurrence can either be in a sentence, in a document or other ways of defining the location of words. The trust we have in the model is that it will cluster words that co-occur and separate words that do not co-occur. In this regard it is important to note that LDA topic model takes 2 important inputs from the researcher(hyper-parameters): the number of clusters(K) and the size of the topics when the model is initialized. This creates a set of challenges, the topic model assumes that we know the number of topics but researcher by definition does not and it assumes (unless a very rarely

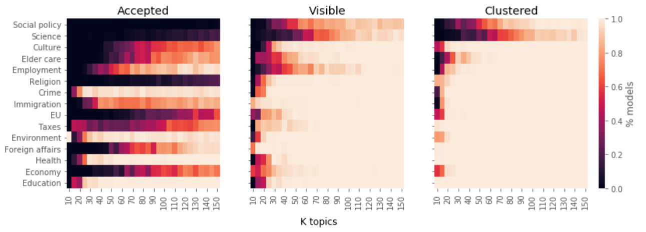

Figure 2. ”Discovery across model runs, locating the optimal Model. A test data set was created by planting 15 categories. Running more than 100.000 models varying K, the number of topics, we report the probability that a model detects the category. Each figure displays the percentage of models detecting it, varying the detection criteria. To the right we define a model to have detected the category if the category is the dominant part of at least one topic, regardless of quality. Here we see that almost all models do indeed locate the planted topics. In the middle a category is detected if a predictive word from the category is present in the top 20 words ordered by the Frex score. To the left we include only accepted topics, i.e. topics not deemed as ’junk’ topics by the researcher. The criteria here is that predictive words from different categories should not be present in the top 20 words as ordered by the Frex score. The last one is the important one, as only accepted topics can be considered discovered. Here many categories are very often overlooked.”

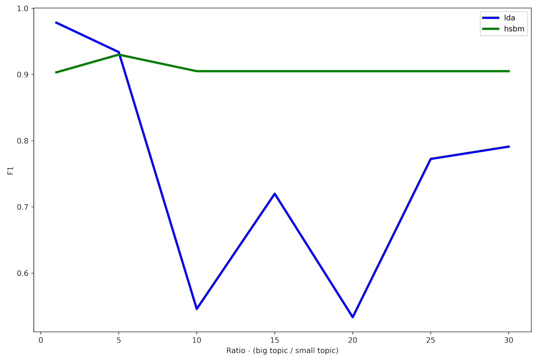

Figure 3. ”HSBM vs LDA: Size Imbalance. Here we test the ability of HSBM and LDA of locating 3 simulated topics while varying the size of one topic (500 documents;500 documents;500 imbalance factor). Performance is measured using the F1 score as the degree to which the discovered topic proportions of each document corresponds to the known simulated topics. As the ratio increase we see LDA collapsing the two smaller topics into one, and splitting the larger into 2. Note that the constant error of the HSBM model corresponds to it locating 3 very small extra topics one for each of the 3 planted topics. As each planted topic is drawn from a zipf distriution, the most frequent word of each of the topic convinced the HSBM model that they should be described by their own topic.”

specified hyper-parameter is manipulated) that topics have approximately the same size (1/k) even if we know that Topics are typically highly unequal in size and similarly to words follow heavy tailed distributions such as the well known Zipf-law for word occurrences (Peixoto (n.d.)).

The challenge of finding the right number of topics is well known (Nelson (2020); DiMaggio et al. (n.d.)). Scholars argue that one can find the appropriate number of clusters by reading the outputted list of words and judge the adequacy of the number of clusters. Too few clusters will return conglomerates of different topics, too many clusters will split single topics into multiple clusters, both these scenarios can be detected by the researcher. This assumes that one can see when a topic is split in two or then 2 topic are combined into one. However, if the topics are different in size, as we expect, then the smaller of two topics clustered into a larger one will not necessarily have any of the most prevalent words in the displayed output. This means that smaller topic can both create measurement error, but also that the researcher is left to conclude that the small topic does not exist in the data.

3 reports on percentage of models that various planted topic is located by the LDA Topic Model. On the horizontal axes we have the different topics and on the vertical axes is the percentage of runs the topic is found. Important to note is that in the simulation the number of topics is varied from 10 to 150 and the total number of model runs is ¿100.000. Several topics, although planted, are seldom found by the model. This points to the problem of letting the model determine the universe of categories pertaining to a specific corpus. If the researcher is lead by the Topic Model output in understanding what politicians talk about and how, then she will be systematically biased towards the specific types of topics the model can locate without knowing the precise qualities of these topics.

The fact that we can’t uncritically trust the model output has further consequences for discovery. Scholars have argued that computational models “...can surprise, challenging presumptions or pre-existing theory, and lead the social analyst to abductively generate new theory by imagining what would be socially required for those patterns to exist”(Evans and Aceves (2016)). However, given that we can’t trust the model output, other scholars have argued that ”Topic models must find what we know is there. Ultimately, a topic model’s trustworthiness must be determined by informed human judgments.”(Ramage et al. 2009, 4). If we differentiate between different degrees of discovery we might say that topic modeling, in the computer lead research procedure, only allows for weak discovery where we can only find the categories we already know, and that our readers know. This is a problem from the point of view of grounded theory which was in part precisely concerned with the development of novel categories from the data partly independent of assumption and prior knowledge of the researcher.

If we can find better algorithms, is it then possible to rely on the co-occurrence assumption? Hierarchical Stochastic Block Models(HSBM) have been demonstrated to perform much better than topic models and they do not require any priors. HSBM are able to find clusters with varying sizes (see Figure 3) and does not merge clusters with little or no overlap. While HSBM are a welcome improvement of unsupervised methods for locating topics in large text corpora this does not necessarily imply that we should stick with the computationally lead analytical strategy. This is at least partly because the models still rely on word features and word co-occurrences, and there is an unknown and uncertain relation between word features and the meaning of a text.

If we return to the famous Robert Firth quote, it is important to note that Firth idea of company was not only other words but what he called the ”context of the situation”. ”Context of the situation” was originally proposed by Malinowski in order to simultaneously broaden the scope of the relevant context and stress that the situation should always be accounted for in the analyses of linguistic expressions. The broader context was for Malinowski both the culture and environment, the ”general condition under which a language is spoken” (Malinowski et al. 1994, 6). The situation captures the occasion, the prior events, the aim and function of the statement that gives the statements it’s meaning (Malinowski et al. 1994, 6). Firth versions of the context of situation is similar with a focus on the persons and personalities, the verbal and non-verbal actions, relevant objects and the effect of the verbal objects

Interpretation and refinement

In the *pattern refinement* step the researcher performances what Nelson calls a *computationally guided deep reading*. This step is supposed to 1) confirm the plausibility, 2) add interpretation and 3) possibly modify the patterns in order to provide an holistic reading. We move from mainly discovering categories to interpreting them. According to Nelson, there are two problems the computational guidance is supposed to solve: the natural limits to scale deep reading and the biased nature of our reading. Both of these concerns are handled by algorithms that identify texts ”that are representative of a particular theme” and that can calculate the relative prevalence of that category (Nelson 2020, 24). A careful application of these algorithms enables the researcher to ”read” and ”interpret any amount of text without the burden of reading the whole text”. Furthermore, both the researcher and the research community can ”trust that then a quote is chosen as an example of something, it is not an outlier but is indeed representative of some theme in the text” (Nelson 2020, 24). From reading the most prevalent words and documents the researcher then labels the topic as being about something specific and based on the reading of the paradigmatic cases writes an analysis of the topic. Saturation is here ensured by the model, that has ”read” the entire corpus and now presents the paradigmatic exemplar to human analysts. Instead of the human having to read through extensive number of cases, as is common in grounded theory, the computer has done so. In other words, while human analysts reads and interprets, it is the computer which leads the selection of documents and is guarantees robustness of the pattern and the representativeness of the document. This assumes that the model has correctly measured and delimited a topic and that words used in text particular meaning match on to the particular meanings that they take on in the other text. Furthermore, it assumes that the reader is able to identify and analyses the categories without substantial exposure.

Even if we assume that the model has found a relevant category and delimited it properly, we can’t assumes that the documents that said to contain a lot of given topic are paradigmatic or representative of a larger topic. It might be that it contains many of the words pertaining to a topic, but that does not entail that it represents the meaning of a given topic in other documents. A paradigmatic case is a case that supports learning, most clearly illustrates, develops and explicates the logic active in many other cases. Locating a paradigmatic case then only makes sense in context of having analysed many other cases which all support taking a given text as the exemplar best able to illustrate the logic of the wider set of cases. This support is not merely word counts but actually meaning-making. Reading a limited set of documents might then make the analysts severely misunderstand the wider logic revealed by a specific case. A second problem is that CGT procedures do not demand that the researcher to become familiar with the field of study. The idea that we can read very little text and get the model to read the rest fails to see that the reader herself has to become qualified to interpret meaning-making in the field. While it has been argued that CG does not replace the competent analysts, the process of gaining familiarity with the field is always external to the proposed research design. Like ethnographers, the researcher has to go through trial and error in order to test their interpretations and come to terms with both what people actually mean and why the researcher’s initial interpretation may have been wrong. Reading a few documents does not support the qualification of the human reader. Even if the researcher is a domain expert the corpus will have certain specificities that has to be learned and accounted for. Furthermore, given that the ambitions is theory development the domain expert can’t, by definition, rely on prior knowledge but has to learn anew - something which requires intensive and extensive reading. In grounded theory the sampling procedure is controlled by the researcher in order to test, develop and further theorise in close relation to the data at hand.

Validation and Measurement

We now turn from discovery and interpretation of patterns to their validation and measurement. How do we validate the topic model before we use it as measurement? In most applied research only inspection of top words and documents, a version of face validity, is used. In the more programmatic statements(grimmer and steward and Nelson), author argue for indirect validation strategies, such as concurrent validity and/or predictive validity. Firstly, it is assumed that face validity and indirect validation measure are sufficient secondly that direct validation is not possible. Here we will present some evidence that challenges these assumption(for a full test of Topic Models for measurement see AUTHOR).

The inspection of top words(most prevalent words in a topic) and documents is *the* way in which a topic is taken to be meaningful. It is assumed that if these top words and document are clearly about a given topic then the model has discovered a valid topic which can be used for measurement. This assumes that the top words and documents are representative of the wider distribution. However, we have no guarantee that top words represent the meaning of unobserved words and documents, because the topic is interpreted and labeled based on a biased and non-random sample. There is certain tension between interpretation and validation. If one takes into account more than the top 10 or 20 words it is hard to interpret, yet labeling a topic based on the top words gives little certainty of the precision of the classification.

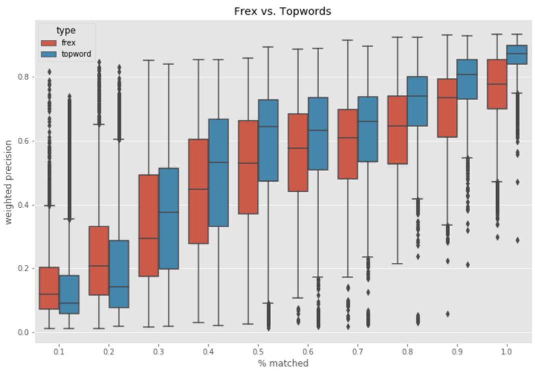

Figure 4. The figure shows the efficacy of the quality inspections from the two different representation schemes. The widely used frex scores (red) as defined by Bischof (n.d.) and the most probable words (blue).

In Figure 4 you see the results of simulation where we have known planted topics with known words associated with the topics(see AUTHOR for a full description). On the horizontal axes we have the precision of the topic classification and on the vertical the percentage of top words match between our known dictionary and the topic model. While a higher percentage of matches does improve the precision of the classification there is still a high variation. With 0.9 top word matches you might still end op with classification that has a 0.60 precision(even as low as 0.20), that is 4 out of 10 documents will be classified wrongly. Top word inspection by no means ensures good measurements.

What about other from of indirect validation, such as concurrent validity and predictive validity? It is important to note that normally these types of indirect validation are used when you have two indicator variables of the same theoretical construct and you want certainty that the indicator variable is measuring the theoretical construct you claim. But this assumes that the indicator variable itself is properly measured, ie has no systematic measurement error. We can't use indirect validation to ensure against measurement error(see AUTHOR).

The obvious solution to this problem is to validate the measure directly by coding a random sample from a given topic and seeing how often the model is right or wrong in its classification. It is standard procedure in supervised machine learning to test the models performance against a test set(for some examples in the social science. While some might argue that the topic is a latent feature that can't be directly validated, it can be validated once the researcher has labeled it a specific topic. While in the analysis of latent dimensions using survey responses it is hard to validate a model by looking at cases within latent dimensions, when working with text it is much more straight foreword because you can read and interpret each individual document(Martin (2018)).

Summing up on our run through of computational grounded theory we believe that some aspects of the procedure makes it less in line with concerns and ambitions within anthropology. LDA topic models can only support weak discovery of what the researcher already knows. Going against, for example, the ontological turn in anthropology commitment to producing novel theories and novel conceptualizations from field work. Likewise idea of minimal immersion does not ensure that the researcher understands the point of view of those studied. Lastly, the assumption that co-occurring words is an appropriate level of analysis goes against anthropological theories of meaning, such as Malinowski idea of the ”context of the situation”. Below we will present a framework which combines computational models and in-depth reading in ways that seeks to allow for strong discovery, support immersion/learning and tries to incoperate Malinowski idea of ”context of the situation” into it's procedure.

From computer lead to computer assisted learning and measurement

In this section we will take the consequence of the above critique and argue for a computer *assisted* approach to text analysis. First, we will reformulate the relationship between computational text analysis and in-depth text analysis, and try to clarify how the different modes of analysis complement one another. Then we will present the outlines to a methodological frameworks which moves from 1) discovery; through 2) interpretation; and to 3) classification and measurement.

Given the critiqued raised above, how should we think of the relation between the qualitative and computational when working with large scale text data? What does it mean to move from computer lead to computer assisted analysis of text data? It first of all means that the last resort of justification of a given interpretation and/or classification of text is to be found in a qualified human reading rather than a computational model. First of all by demonstrating that the researcher understands how signs are used within a specific community(or context) and secondly how an given interpretation adds to, or is line with theoretical interests of an community of inquiry. Part of this work relies on an overview of what is going on within a field and partly on saturation, namely having seen enough cases, containing enough variation, to arrive at robust understanding and theory of the focal phenomena.

Both gaining an overview and reaching saturation is typically a challenge. Overview is hard given the large heterogeneity in what a corpus is about and the high variations in size that different categories typically have. Saturation around any give category is hard because the category will very likely be very rare in the data and hence locating these cases can't be archived through random sampling. They typically do not have a known location and hence can't be found by looking somewhere specific. Even within each category the researcher needs to gain an overview and saturated understanding of each relevant subcategory - here again we are likely to observe high variation in size between the different subcategories. We also do not simply want to learn from cases but also measure their prevalence or relations to other variables, and hence need to classify large amounts of texts.

A important constraint on our framework is that it starts from the premise that we do not known how words, word relations or the context of the situation in any general

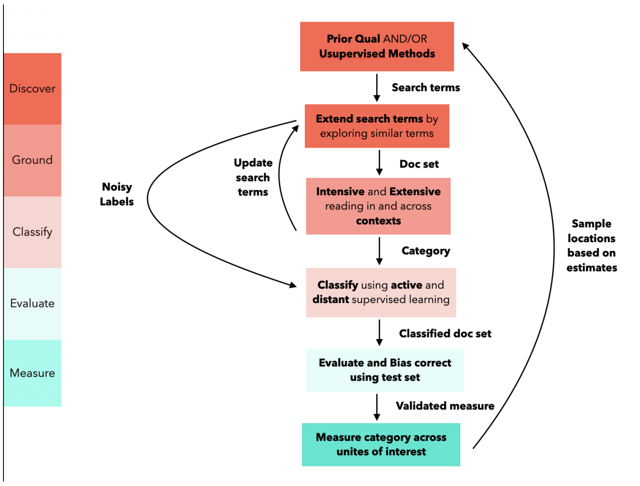
Discover
Ground
Classify
Evaluate
Measure
Prior Qual AND/OR Usupervised Methods
Search terms
Extend search terms by exploring similar terms
Doc set
Intensive and Extensive reading in and across contexts
Category
Classify using active and distant supervised learning
Classified doc set
Evaluate and Bias correct using test set
Validated measure
Measure category across unites of interest
Noisy Labels
Update search terms
Sample locations based on estimates

Figure 5. The CALM workflow.

sense map on to meaning making. All we can hope for is that certain features(word, word-relations, author, audience, recipient, time and place) might work as effective indexes of certain meanings, as opposed to being neccesarily related to it. Thereby, we take account of the contextuality of meaning that both Malinowski and Firth staged with the idea of the context of situation, but we not mean to contribute to specifying in general terms how situational elements relates to meaning. The meaning baring unite is the *sentence in context*, jet precisely what elements of this context which are relevant to the meaning of a sentence is something learned within the specific setting that one is analyzing. This happens through manual readings of text and active seeking out relevant contextual elements. Mainly through this type of reading can one gain robust idea of whether a sentence is of a certain general type and what we might learn from it.

This means that manual reading and classification is the practice through which discoveries are made(although a computational model does assist in finding cases), interpretations are grounded and proper classification is ensured. A qualified reader and their analysis is the ”ground truth”, yet computational models ensures that the human reader finds enough cases with enough variation to ensure saturation and which enables the scaling of the classification made by researcher. Our framework is quite conservative in its division of labour between the machine and the anthropologist.

Figure 5 is a simple version of the CALM workflow. It goes from initial discovery, grounding interpretation, classification, validation and measurement. Like the computer lead workflow there are various stages where one moves back and forth. Below we will detail these steps and try to explicate the logic.

Discovery

In CALM framework the goal of discovery is two fold. First, discovery might propose a relevant partition and categories by clustering certain words. This clusters might support, supplement or challenge those the researcher already has from prior qualitative work, domain knowledge or theory. Second, it provides the researcher with words that can be used as an initial search set, seeds, used in order to find instances relevant to further ones interpretations of a give category.

Before exploring partitions of the data it worth while to consider what structural features ones category of interest might have or has been found to have. There are other patterns then the ones captured by a bag of word representation that might be useful. It might be that the type of content one is interested in is best located by also keeping some of the sequential information in the sentence. Like others we have been interested in topics and here co-occurrence might be sufficient information for the model. We have found the HSBM to provide high quality clusters and have used this model for initial discovery of relevant partitions and search terms. The fact that it can find small topics is very important because it is typically the smaller categories that are of interest and that are hard to locate by other means.

The particular clusters that the model returns is supportive if it finds the same categories that our theory/prior work would lead you to expect. It might also supplement our theoretical expectations by proposing new categories that we had not thought about. Lastly, it might challenge our theoretical expectations by redrawing the boundaries around an set of known categories. An important difference from computer lead approach is that in this phase everything is taken to be at the most uncertain level of inference. The words that are clustered together by the model are conceived as highly uncertain sign of a given category. The topics that the model does not find are not necessarily then deemed irrelevant and neither do we necessarily respect the boundaries proposed by the model.

The search terms provided by HSBM are extended through searching for similar words. We use Word2vec to get a measure of similarity between words. Word2vec is what is called a sallow deep learning model where the words loading’s on a specified set of dimensions is tuned so that is able to solve a task of predictiing neighboring words (Mikolov et al. (2013)). Word2vec has proven to be an effective way to expand the dictionaries (Tulkens et al. (2016)). Just as with HSBM we use the word2vec to find what we expect to be effective search terms and not provide us with cultural logics within a corpus(for this more ambitious and risky endeavour see Kozlowski et al. (2019)). The task is similar to a snowball crawl where one transverses a given area in a network in order to sample cases. In order not to get stuck in particular non-productive areas of the embedded space we either purposefully and automatically alter the search term list. This ensures a greater search coverage.

In this phase the researcher is analyzing on the level of words. Word and word relations can be effective in stimulating our imagination regarding the actual meaning making within each document. Discovery here is one of possibilities, where the theoretical imagination is sparked to life by patterns of words. The end product is a set of categories with a connected set of search terms. Each category can naturally have subcategories with connected subset of search terms. None of which are taken to be empirically grounded jet.

Interpretation and grounding

In the "interpretation and grounding step" the researcher retrieves a documents set by applying the search terms. The goal is to first to judge the adequacy of the category, given the documents. Second to read and learn from an extensive and varied set of documents relevant to a given category(qualifying the researcher). Third, to write an extensive analysis of the category, it's relation to the social process under study and it's important dimensions of variation(justifying the category and the interpretation of various cases). Fourth, to construct a coding scheme with a definition of the category, illustrated with exemplars and boundary cases, that can be used to classify content in transparent, interpretatively valid and reliable manner. Lastly, it is also here that the researcher becomes certain of the most theoretically productive analysis given data.

The search terms are meant to find the relevant instances, to overcome the problem of rarity and secondly ensure enough within-category variation to support theoretical saturation. The analyst can apply the constant comparative methods to the document set in order to develop defining features, its subdivision, continuum's and analyse its relation to other categories (Glaser 1965, 439). There is an important trade-off in the construction of the search terms. An too inclusive strategy will reproduce the problem of rare events where too many irrelevant documents are retrieved - the problem of low precision. On the other hand we have the problem of low recall where the search set is too restrictive so that we miss many of the relevant instances of the phenomenon. Importantly recall and precision should is here relative to insuring saturation, not measurement. A solution to this problem is sequential sampling(Mario Smalls) where the search terms list is constructed on the basis of the developing the theory, and hence moving from a more restrictive to a more inclusive search set, or going into depth with specific subcategories.

In this mode of reading an important question is the organization and construction of the text data - similar to what is call unitarization in content analysis(Krippendorff (2004)). Our own work is primarily with Facebook data and some of the tricky issues with unitarization of documents are solved by the fact that Facebook itself is organized in relatively comprehensible and manageable unites. Typically, post are on a group, a page or personal profile and with connected comments and comment-to-comments. Yet, there is still the question of how much context should be surrounding each post and comments, when the researcher is reading. This has to be determined partly by theoretical focus and partly by experience. While, as argued above, we are never sure what each encounter demands of us in terms of context tracing it is for practical reasons worth while trying to organizes ones material in such a way that relevant context, on average, is displayed or easily accessible when reading. For example in analysing contention refugee solidarity groups on Facebook, each post or comment was analyzed together with the post and other comment connected to it in order to use the information from surrounding text to infer what a comment or post was about and if it indicated opposition against third parties (Author).

This is the step that both ensures the interpretative accountability and validity of the text analysis and its theoretical relevance. This implies that not only coding and memo writing is necessary but also accounts of the development of ones own interpretative journey moving from one understanding to the next, documenting the events where one reached a new understanding of the meaning-making within the field, this is what Lichterman calls interpretative reflexivity (Lichterman (2017)). The steps ends with both an qualitative analysis of category, but also a more concise coding scheme that in principle can be used by others.

Classification and Validity

At this point in the analysis we should have a qualified researcher, a developed understanding and analysis of a set of categories and an applicable and valid coding scheme. The next step is to apply this coding scheme to classify content in order to both train a machine learning model and test it.

At this stage the coding process is closed and no theoretical adjustments to the category is to be made. This does not mean that there is not any interpretation going on, but this is at the level of correctly apply the code to content rather than a correct understanding of the category. Therefore, the prior qualification of the reader is important because there still a lot of non-explicated knowledge that might be relevant for the application of the code. Ideally the coding scheme is for externalizing, objectifying and making transparent the knowledge gained from in-depth reading. Yet, while before the coding necessarily involved research assistance in order to scale the classification, with the advent supervised machine learning it is possible to rely on much less coded data.

For training the machine learning model, the CALM framework relies on active learning procedures(for a social science example see Wiedemann (2019)) where the machine learning model uses model uncertainty in order to find the most informative cases to be coded. In traditional supervised machine learning the training data is given to the model all at once and hence risks both a lot of redundant information about certain cases and lack knowledge of others. For example in our own prior work we used nearly 9000 examples for training a model on a single category and 3000 for testing. Therefore, supervised machine learning has traditionally been seen as a very expensive compared to unsupervised methods for classification. However, with the rise of active learning and transfer learning this cost is radically declining.

Transfer learning has dramatically changed the efficiency of the features used to solve classifications tasks. Language models that have trained on huge amounts of text have been shown to enable what has been called few-shot learning, which means that the model only needs very few training examples in order to achieve high performance (see for example Devlin et al. (2018)). A second set of features is the extensive dictionaries developed in the prior steps. These are used as noisy labels, which means that they are assumed to contain a lot of noise and are only seen as indicative of category in so far that they do not contradict the human label.

The last aspect of classification is constructing the test set. While the training set becomes smaller due to more efficient models, the same is not the case with training. As we have argue automated text classification can be highly biased and with the use of language models trained on huge data sets we are opening op to the biases that are known to exist in these models. While a lot of these biases should be handled in training this is not certain. Besides the biases from the language models many other source of bias might occur: word usage might differ across population making it easier for the model to correctly classify from some parts of the population, there might be more data from some periods and more. Again this points to the importance of the test set and that it is large enough to analyse the performance of across relevant units of analysis in order to ensure that ones results are not drive by measurement error. This is also the Achilles heal of automated text classification, because estimating the performance of the model across the relevant variables quickly demands a unfeasible very large test set especially when the categories are rare(rarity increases sample size in order to ensure statistical power).

Discussion

In this discussion we discussion the CALM frameworks relation to recent focus on abductive inference, a trend within both qualitative sociology and anthropology, and the concerns with reproducibility and credibility central to computational grounded theory.

Given that grounded theory was about theory generation what implication do recent critiques and alternatives focusing on abduction do to our discussion? A stronger focus on existing theories and working with multiple theories, rather then assuming that emerges out of the data, fits our framework well. Precisely, when the idea of natural clusters is left behind and the search is lead by the researcher then it is possible to try out different theories and make risky abductive inferences. Abduction, to be effective, also dependence upon robust evidence or else the suppressing fact that should provoke creative theorizing can't be trusted. From a substantive, and not methodological point of view, the most boring answer to a suppressing fact is that it results from measurement error. While it is true that certain partitions returned by the model might lead to creative hypothesis generation, abductive analysis as layout by Tarvory and Timmermans also rely on extensive empirical work following the initial abductive inference in order to arrive at well constructed and workout theoretical account(ref). CALM supports this through extensive sampling across different axes of variation. This furthermore, also points to CALM focus on immersion and the "context of the situation" so that the researcher both in general learn sign usage within a given context and able to classify content correctly.

One of the central concerns of computational grounded theory is reproducibility and the CALM procedure seems less reproducible. It is clear that with extensive readings and a variety of search operations it is a more complicated procedure then running a few topic models varying on number of topics that should be returned. On the other hand every search operation can be traced, documented and reproduced if the operations are coded in a programming language. Most importantly reproducibility from a qualitative point of view is more about understanding the perspective and experiences that created a certain interpretation. Here extensive reading, quotes and explicit theoretical choices in the analysis of a category provide a more solid background than reference to model output.

The shift to the researcher and the discrediting of the computational model also raises the question of credibility. The credibility of computational grounded theory to great extent relied on trusting the model, which we argue one should not. How then is the credibility of the qualitative analysis communicated and ensured. Again we would partly appeal to classic qualitative strategies, such as demonstrating the capacity of the human reader, through their display of sensitivity to subtle differences in meaning making and their ability to describe the specifics around their field site. Secondly, the extensive search in large text data sets makes it possible to ensure within case generalization - something qualitative studies have been critiqued for lacking. The logic being that given the map of the corpus that the models provide and the places that have been searched out by the researcher there is good reasons to believe that all relevant parts of the corpus have been analysed.

Conclusion

In this article we argue against the computer lead version of computational grounded theory. We argue that this approach puts unwarranted trust in the models ability to locate the relevant categories in the data. Firstly, we demonstrate that the LDA topic model used extensively in the social sciences can't locate plated topics, and tends to create conglomerates cluster made up of multiple topics that can't be detected by inspecting the model. Secondly, we argue that we can't assume to know how meaning maps on to word patterns and hence can't rely on any given model to provide os with relevant categories in the data. Secondly, that Computational Grounded Theory idea that the model can locate the few representative documents fails to support the qualification of researcher. Lastly, we demonstrate the validation strategies used in contemporary social science, face validity and indirect validity, does not ensure against substantial measurement error.

As an alternative we argue for an computationally assisted approaches that puts the researcher in charge of all operations but uses the models ability to locate potentially use full patterns and word similarities. This approach builds extensive search term lists with the help of computational models, using the lists to retrieve documents set that are read and analysed in-depth using the constant comparative method. In line with Grounded Theory the research process in its discovery and interpretation phase is build up around theoretical sampling where the researcher is constantly sampling in order to support theoretical saturation around a given category. The phase of interpretation support the qualification of the researcher and an extensive analysis of a given category, which is turned into coding scheme used for classification. The classification step uses a active learning framework combined with transfer learning in order to as effectively as possible train a model to replicate the researchers classification. The costs saved in training the model are instead used to test the model and investigate differential biases across the variables of interest.

References

Baumer, E. P., Mimno, D., Guha, S., Quan, E. and Gay, G. K. (2017), ‘Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?’, *Journal of the Association for Information Science and Technology* **68**(6), 1397–1410.

Biernacki, R. (2012), *Reinventing evidence in social inquiry: Decoding facts and variables*, Springer.

Bischof, J. M. (n.d.), ‘Improving and evaluating topic models and other models of text AU - airoldi, edoardo m.’, **111**(516), 1381–1403.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018), ‘Bert: Pre-training of deep bidirectional transformers for language understanding’, *arXiv preprint arXiv:1810.04805* .

DiMaggio, P. (2015), ‘Adapting computational text analysis to social science (and vice versa)’, *Big Data & Society* **2**(2), 2053951715602908.

DiMaggio, P., Nag, M. and Blei, D. (n.d.), ‘Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding’, **41**(6), 570–606.
URL: *http://www.sciencedirect.com/science/article/pii/S0304422X13000661*

Evans, J. A. and Aceves, P. (2016), ‘Machine translation: Mining text for social theory’, *Annual Review of Sociology* **42**, 21–50.

Glaser, B. G. (1965), ‘The constant comparative method of qualitative analysis’, *Social problems* **12**(4), 436–445.

Glaser, B. G. (n.d.), ‘The constant comparative method of qualitative analysis’, **12**(4), 436–445.
URL: *https://academic.oup.com/socpro/article-lookup/doi/10.2307/798843*

Glaser, B. G. and Strauss, A. L. (2009), *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Transaction Publishers.

Grimmer, J. and Stewart, B. M. (2013), ‘Text as data: The promise and pitfalls of automatic content analysis methods for political texts’, *Political analysis* **21**(3), 267–297.

Holbraad, M. and Pedersen, M. A. (2017), *The ontological turn: an anthropological exposition*, Cambridge University Press.

Kockelman, P. (2017), *The Art of Interpretation in the Age of Computation*, Oxford University Press.

Kozlowski, A. C., Taddy, M. and Evans, J. A. (2019), ‘The geometry of culture: Analyzing the meanings of class through word embeddings’, *American Sociological Review* **84**(5), 905–949.

Krippendorff, K. (2004), *Content analysis: An introduction to its methodology*, Sage publications.

Lee, M. and Martin, J. L. (2015), ‘Coding, counting and cultural cartography’, *American Journal of Cultural Sociology* **3**(1), 1–33.

Lichterman, P. (2017), ‘Interpretive reflexivity in ethnography’, *Ethnography* **18**(1), 35–45.

Malinowski, B. (2002), *Argonauts of the Western Pacific: An account of native enterprise and adventure in the archipelagoes of Melanesian New Guinea*, Routledge.

Malinowski, B. et al. (1994), ‘The problem of meaning in primitive languages’, *Language and literacy in social practice: A reader* pp. 1–10.

Martin, J. L. (2018), *Thinking through statistics*, University of Chicago Press.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), ‘Efficient estimation of word representations in vector space’, *arXiv preprint arXiv:1301.3781* .

Nelson, L. K. (2020), ‘Computational grounded theory: A methodological framework’, *Sociological Methods & Research* **49**(1), 3–42.

Nelson, L. K., Burk, D., Knudsen, M. and McCall, L. (2018), ‘The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods’, *Sociological Methods & Research* p. 0049124118769114.

Peixoto, T. P. (n.d.), ‘Hierarchical block structures and high-resolution model selection in large networks’, **4**(1), 011047.
URL: *https://link.aps.org/doi/10.1103/PhysRevX.4.011047*

Ramage, D., Rosen, E., Chuang, J., Manning, C. D. and McFarland, D. A. (2009), Topic modeling for the social sciences, *in* ‘NIPS 2009 workshop on applications for topic models: text and beyond’, Vol. 5, p. 27.

Small, M. L. (2009), ‘How many cases do i need?’ on science and the logic of case selection in field-based research’, *Ethnography* **10**(1), 5–38.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B. and Daelemans, W. (2016), ‘A dictionary-based approach to racism detection in dutch social media’, *arXiv preprint arXiv:1608.08738* .

Wiedemann, G. (2019), ‘Proportional classification revisited: Automatic content analysis of political manifestos using active learning’, *Social Science Computer Review* **37**(2), 135–159.