# Crowston, K., Allen, E. E., & Heckman, R. (2012). Using natural language processing technology for qualitative data analysis. International Journal of Social Research Methodology, 15(6), 523-543.

*Great to introduce concepts that I will use. (existing qual methods, existing NLP based apps and potential issues from using NLP in qualitative research)*

**Introduce existing problem and techniques**

However, because such data are textual, they require considerable manual effort to analyze. This work is tedious and difficult for humans to do reliably at scale. As a result, qualitative research addressing important questions in social research often relies on small sample sizes because of the analysis effort required

If successful, NLP tools could advance the work of social researchers by extending the capabilities of current tools, and enabling researchers to explore massive data sets at a greater depth

Content analysis is a qualitative research technique for finding evidence of concepts of interest using text as raw data (Myers, 1997). The result of the coding process is a text annotated (or tagged) with codes for the concepts exhibited (Miles & Huberman, 1994). In the approach we describe, codes are applied based on the features of specific segments of text, rather than with the goal of understanding and interpreting the entire text as a whole. The goal of coding texts is to be able to study the relationship between concepts as expressed in the text.

A key concern in coding is reliability, as measured by the degree of inter-rater agreement (also known in some circles as inter-coder reliability), that is, whether different human coders working on the same text identify the same set of codes. If coders do not agree, then they discuss the coding until they reach a better level of shared understanding of the code. Codes and coding decisions are documented in a codebook, which includes definitions of codes, and best examples of when to use them. However, a great deal of tacit knowledge is often used in coding, meaning that coders need to be trained extensively to code reliably. Once the coders are coding reliably, they must read the texts to code them for the concepts of interest, which can be quite labor intensive for a large corpus. Furthermore, the tedium of the work makes it difficult for human coders to do reliably. Being able to analyze larger data sets is necessary to examine and compare multiple groups or to identify smaller effects within a group, but research teams often face limitations in the scope of analysis based on the available work force. It is this problem of reliable coding at scale that we seek to address by using NLP technology.

**Recap of NLP**

NLP is a computational approach to text analysis (Jurafsky & Martin, 2009). It ‘is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications’ (Liddy, 2003). NLP is a large area of research and application, with numerous research problems and approaches. Example NLP tasks include automatic text summarization, machine translation, information search, and question answering. In the current paper, we discuss how NLP techniques can be used to automate (fully or partially) the process of qualitative data analysis by identifying segments of text that provide evidence for concepts of theoretical interest (i.e. coding) (Evans (1996) gives an overview of other computer applications).

On the other hand, corpus-based statistical methods apply mathematical techniques to build models of linguistic phenomena from actual examples (Manning & Schütze, 1999). Such an approach can be quite successful given a sufficiently large set of training examples and indeed has become the dominant mode of analysis for large corpora.

Existing apps: Atlas.ti, Hyper-research or Nudist (lower level language processing) vs Linguistic

Inquiry and Word Count program (<http://www.liwc.net/>)

**Results and discussion**

In manually coding the data, the researchers chose the thematic unit as the unit of coding, a common choice in qualitative data analysis. A thematic unit may consist of a word, phrase, sentence, or even an entire paragraph that is felt by the coders to provide evidence of a given concept. Unfortunately, with this variability in scope, it is difficult to exactly match the boundaries of text to be captured using NLP rules. For the results reported above, any overlap between the text coded manually and that coded automatically was considered to be a match, as is often done when comparing human coders. To facilitate future comparisons of human and automatic coding though, it would be better to pick a less amorphous unit of analysis for coding, such as a sentence or even an entire message.

Finally, an important component to consider in the evaluation of NLP-enabled content analysis is the potential cost-benefit to a research project. While NLP can potentially automate parts of the coding process, additional effort is required to develop and validate a rule set. In the approach we took, NLP coding was built drawing upon a manually crafted qualitative coding codebook (as is required for any such qualitative study). However, the NLP rule set required additional development for the rules and testing time to assess performance, both requiring the time of a trained NLP analyst. For a large-scale analysis system handling very large volumes of textual data, particularly discourses spanning long periods, some development time might also be needed to adjust for changes in data format, new discoveries, and evolution in both the data content and the analytic thinking.

In light of the additional work needed, an NLP-supported approach would not make sense for small (e.g. a thousand or so messages) unique data sets that can be handled by training content coders within a relatively short time span. Furthermore, we note that the NLP approach is only appropriate for theoretical concepts that find a regular expression in text. With the current approach, NLP would be unlikely to work for coding that draws heavily on subjective interpretation and context. However, for suitable codes, and after development resources have been invested, benefits can be realized for large-scale studies by processing and analyzing large volumes of data with reduced human coder effort. Specifically, the investment of time in writing rules should enable order of magnitude reductions in the effort needed to code additional text, potentially allowing the analysis of hundreds of groups with hundreds of thousands of messages. Indeed, such automation is arguably the only way to reliably handle such large amount of text that would otherwise require hundreds of person-years of coder effort.

# Ruelens, A. (2022). Analyzing user-generated content using natural language processing: a case study of public satisfaction with healthcare systems. Journal of Computational Social Science, 5(1), 731-749.

*More suited as an example where they used NLP for research. As well as some tools examples and steps taken in analysis*

**METHODS**

Previous research of online readers’ comments has often relied on **manual analysis, which is a limited factor in the ability to analyze larger corpora of comments**. In this study, we use automated natural language processing (NLP) tools that facilitate this type of textual analysis. **The two main text mining methods applied in this study are word frequency analysis and sentiment analysis**. Both methods are extensively used in the studies based on computational approach (e.g. [24, 39, 40]). The frst method, word frequency analysis, as its name suggests, involves **identifying most frequently used terms in the body of user comments**. This technique is based on the idea that words which are most frequently used by the commenter indicate issues of higher importance to the user [35]. This method is inductive in identifying the topics of relevance, compared, for example, to the predefned structure of the survey instrument. Given the focus of this study, we were particularly interested in the most common terms related to healthcare and medical services. In the second part of our analysis, we employed text **sentiment analysis also known as emotional polarity computation. The main aim of sentiment analysis is to determine the subjective opinions of online users with respect to a specifc topic** (for an overview of sentiment analysis see [31]. These opinions can be of an evaluative or of a judgmental nature. Moreover, they can also refect the emotional state of the user, revealed either intentionally or unintentionally. For the purposes of this study, higher positive sentiment score refects a more positive attitude towards healthcare systems. Although sentiment analysis is a mainstream tool in text mining, this method continues to evolve and achieves greater accuracy and validity.

**RESULTS AND DISSCUSION**

We should acknowledge certain limitations of this study that ofer several opportunities for future work. First, like with much of the research using user-generated content, we should be cautious in generalizing our results. The majority of the individuals who participated in commenting online represent a general NYT readership. As mentioned above, in terms of the socio-demographics, this group of commenters is characterized by medium to high level of education and income

# Parks, L., & Peters, W. (2022). Natural Language Processing in Mixed-methods Text Analysis: A Workflow Approach. International Journal of Social Research Methodology, 1-13.

*Also good for introducing concepts but mostly cites the Crwoston 2012 paper to do so. Proposes an iterative process on how to incorporate NLP to qual analysis*

**Talking about mixing methods (qualitative plus NLP)**

Before situating NLP in the broader literature on mixed-methods text analysis and describing the illustrative example, some words on our approach to mixed-methods within a workflow are in order. The basis for our claim about the advantages of flexible workflow approaches to mixing methods is rooted in scholarly debates on mixed methods, including questions of how to mix methods with due attention to research paradigm compatibility. It also responds to calls to move beyond methods triangulation for corroborating or challenging findings and increasing their validity, or producing different views of the same data (Hammersley, 2008), towards ‘a multimethod logic’ that can blend ‘the sensibilities of both computational and interpretive approaches to social analysis’ (Espinoza-Kulick, 2020, p. 51). Scholars have also called for more attention to be paid to how we engage with computers and analytical software, by unpacking the ontological and epistemological assumptions of research projects and/or computer software (e.g. Jacobs & Tschötschel, 2019) and by leaving space for critical and self-conscious engagement with computers (Hitchcock, 2013), research questions, data and findings, and indeed between social and computer scientists.

**Traditional qual analysis**

Many traditional scholarly approaches in the social sciences involve close reading activities relying on manual analysis. This is the foundation of, for example, formal legal analytical approaches (black letter law or doctrinal analysis) as well as much qualitative content analysis in sociology and political science – whether frame analysis, claims analysis, discourse analysis or other types (Boréus & Bergström, 2017). Depending on their research agenda, scholars tend to take either a predominantly deductive approach, investigating the textual attestation of pre-formed hypotheses, or a predominantly inductive approach, working empirically towards the formulation of one or more hypotheses on the basis of evidence arising from the textual material. The results from these close, manual analyses are rich and detailed, and of vital importance for new areas of study where little is yet known. Content analysis methods are also closely linked to research questions at the ontological and epistemological levels, and crucial for testing and developing theory. Their advantages for these goals are clear and undisputed in terms of the granularity of their results and operationalizability

Nevertheless, it is also widely acknowledged that the limited quantity of textual material it is possible to interpret using qualitative, manual techniques can create a significant bottleneck for an exhaustive scholarly understanding of the content of the textual source material and the domain it stems from. Close, qualitative textual analysis provides rich and detailed information, but is time consuming and can be ‘tedious’, which creates significant challenges in terms of researcher errors (Crowston et al., 2012) and unintentional cognitive bias (Boréus & Bergström, 2017; Crowston et al., 2012). The time required for manual analysis means it is difficult or impossible to apply to large amounts of data (Chakrabarti & Frye, 2017). Sampling is necessary, which can limit the generalisability of results, at least in the sense of the term generally applied in positivist paradigms, often demanded of qualitative methods despite the diverse reasoning behind them (Smith, 2018). In addition to this, the plasticity of content analysis methods can affect the perceived validity and replicability of the research. Though these concepts are not applicable to qualitative manual analysis in the same ways as quantitative analyses, if researchers fail to fully explain their methods and decisions taken in their analyses, results can be challenged. This also makes results difficult to replicate or verify through applications to other data (e.g. Chakrabarti & Frye, 2017; Jacobs & Tschötschel, 2019).

# Nti, I. K., Quarcoo, J. A., Aning, J., & Fosu, G. K. (2022). A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. Big Data Mining and Analytics, 5(2), 81-97.

*Basically using ML to make literature reviews easier. Automated filtering and discovery of papers based on keywords to determine which ML and BDA methods are the most prominent currently (wow so meta)*