MDPI

Article

Text Mining for Processing Interview Data in Computational Social Science

Jussi Karlgren [1,*], Renee Li [2] and Eva M Meyersson Milgrom [3]

[1] Gavagai, Stockholm
[2] Stanford University
* Correspondence: jussi@gavagai.io

Version July 31, 2019 submitted to Multimodal Technologies and Interact.

Abstract: This paper describes how text analysis technology can be applied to the analysis of semi-structured questionnaire responses in social sciences. In unstructured and explorative studies, open semi-structured or even unstructured responses from interviews or observations are an established and useful data acquisition method. The general challenge of questionnaire studies is to analyse the data systematically in order to verify or refute some hypotheses of interest. This is where text analysis technologies can help: they are essentially scoring mechanisms geared towards topical classification of texts or text segments. Using such existing mechanisms poses a specific challenge for studies being conducted today, and experiences from doing so provides a road map for developers of coming systems in the near future. These challenges are illustrated with examples taken from a large cross-cultural study on intimacy currently under way, analysed using a tool which is developed for the analysis of market research data: a task with obvious similarities to processing data for the social sciences. We found that topical clustering and terminological enrichment provides for convenient exploration of the responses, and tabulating presence of observed concepts allows for correlation analysis and inference of relations between variables. This makes rapid hypothesis testing and comparison with e.g. demographic or other set variables possible. We also found that many potential variables of interest were challenging or impossible to extract but that there are obvious paths to tailor current language technology mechanisms to this type of data.

Keywords: Text analysis; Computational social science; Questionnaire processing; Open answers

1. Computational social science and text analysis

Computational social science, as one of its central methods, gathers and processes textual data. These data can be obtained through archive studies, media analyses, observational studies, interviews, questionnaires, and many other ways. Computational text analysis has a broad palette of tools to create structured knowledge from human language. The field of language technology, the basis for text analysis tools, has devoted most of its attention to *topical analysis* of text, to establish what a text or a segment of text is *about*. This reflects the uses that text analysis has been put to: mostly topical tasks, where relevance and timeliness of the information are the priority. The knowledge gathered from texts in computational social science are, in contrast with the general case, only partly topical in nature, and existing text analysis tools and methods are not always well suited to these types of task. This paper will examine how text analysis can be of use, and point out some cases where current technology could be improved to be more effective.

2. Questionnaires and quantifying textual data

There are multiple well-motivated reasons to gather human data through interviews or questionnaires where respondents can use their own language to respond to queries rather than e.g. multiple choice or graded agreement responses.

One reason is to make the data gathering situation less formal and more personal and thus encourage the respondent to provide richer data. Gathering data through natural human language allows respondents to express what they *feel*, *perceive*, *believe* and *value*, in a language they are comfortable with. This allows the analysis to detect the attitude of the respondent to the topic they are responding to. In addition, through an analysis of the language used through the entire interview, general observations about the stance and emotional perspective of the respondent can be made.

Another reason is to allow the respondent more leeway to formulate their responses without too much imposition from a pre-compiled response structure and thus enable the respondent to provide unexpected data or unexpected connections or dependencies between items under consideration [1, e.g.]. This is especially true if the study concerns (1) a vague or indefinite subject, (2) treats matter for which there is no established vocabulary and phraseology, or (3)which is sensitive in some way [2]. An interview situation will also allow dynamic follow-up questions which allow specification of aspects that might not otherwise have been detected and will allow the interviewer to probe for bias in the form of tacit assumptions the interviewee might have about what interests the interviewer [3].

For the above reasons, the design of interview studies and analysis of interview data is an area of methodological debate and research within the social sciences [4–6].

While open answers in surveys and questionnaires provide richer data and reduce the effort and difficulty of formulating the questions in exact form before the fact, they move the effort to after-the-fact-processing of the collected data to get useful results. Open answers are a challenge for analysts: reporting the collected responses together with more quantitative data elicited from respondents is not obvious. Coding procedures—converting open responses into structured form—require time and expertise on the part of the analyst, both of which come at a cost. The effort involved in coding open answers is simultaneously intellectually non-trivial and demanding, but still monotonous: analyst fatigue and frustration risks leading to both between-analyst and within-analyst inconsistencies over time in reporting.[1]

3. Text mining: turning text to quantifiable information

Language technology has over the last decades developed a series of analysis methods and tools for processing factual content in text, mostly for news material, legal documents, technical matters, or other related application domains. Most text analysis methods generalise well to other types of content and other genres: the focus has been on the specific and the topical rather than on understanding background tenor, perspective, or stance.

Text mining is the general term for the systematic extraction of (somewhat) structured insights from unstructured text material using text analysis methods of various kinds. Text mining may refer to a range of tools, built on simple methods such as keyword extraction to more complex language understanding mechanisms, and to a range of activities, from straightforward general text classification to more specific analyses tailored for some application task.

Text mining as an application area traces its roots to the very first steps of language technology, the Linguistic String Project, and *information extraction*, which finds and tabulates pre-identified and carefully formulated structural patterns in text [9–13]. Text mining tools have spread to many practical

[1] E.g. O'Cathain and Thomas [7] and many others; It takes about 1 minute for a human to categorise an abstract, shown by e.g. Macskassy et al. [8] when the categories are already given , If the task is to explore a set of responses and define and revise categories or labels as you go it will involve more effort and require more time per item.

fields in recent years. Overviews of text mining are many; a technical perspective is given e.g. by Aggarwal and Zhai [14].

The more recent generation of text analysis tools score textual material quantitatively in order to then classify the texts into some set of categories. These categories are typically given to the system by a set of manually categorised examples or are derived from the text collection itself. The most popular approach currently is the family of methods known as *topic models* which is generally understood to refer to the specific strand of probabilistic models originally defined by Blei, [15, e.g.]. On a more general level, any procedure which relates texts or text segments to a set of topics based on term usage can be called a topic model, irrespective of algorithmic details.

Topic models have lately been used in e.g. digital humanities for mining historical archives and in media monitoring for mining news feeds and the like. This introduction of new digital methodology for scholarship has not been uncontroversial [16–20, e.g.]. The debate over how to best use new technologies is lively and goes to the roots of what the ultimate research goals of the humanities and the social sciences are. The humanities and the social sciences do not only have different methods than engineering and the natural sciences do, but their goals and aims when they produce knowledge are different, and they approach information differently. However, to some extent, this debate has a technical question in the background: how can the technologies that have been developed for some task be transferred to be useful for another?

4. How text mining can be used in the social sciences

Data analysis in questionnaire studies involves analysis of textual data in order to verify or refute some hypotheses of interest, or to explore a domain to establish hypotheses for continued study. For studies in social sciences, concepts mentioned in texts are operationalised to be clear and invariant. Matching linguistic expressions to the concepts under study reliably and consistently is the main challenge. This lays open every challenge of text analysis: linguistic variation crosstabulated with the interplay between topic, attitude, and respondent characteristics. The degrees of freedom are obviously larger for explorative studies where hypotheses are yet unformed and the concepts are not defined, and in such studies, the analysis must take care not to generalise over potentially interesting variables of interest.

Where today such exploration is based on human coding of textual material, text analysis technology can improve consistency and productivity, if the observable features it makes use of are relevant to the task at hand. As seen above, text analysis is a general scoring mechanism which in most applications today is optimised for the topical classification of texts or text segments. Using such tools poses a specific challenge for studies being conducted today, and experiences from doing so provides a road map for developers of coming systems in the near future.

For research purposes one must be able to revisit the data and reproduce the results using the methodologies originally used for analysis; other research groups will want to be able to replicate the results based on the description given in the original report, and to generalise from the given study to new populations. This poses strong requirements on the *transparency* of methods and tools used. Traditionally, data for social sciences has been coded by human effort. This has been accepted to be transparent, consistent, and replicable, although, arguably, it is so only to a degree.

Recent approaches to text analysis rely importantly on learning and adapting to data; on transfer learning, which relies on applying models learnt on vast amounts of background data; and on end-to-end training, where a model is trained by showing it some examples of a desired analysis and then unleashed on an unencoded data set. This is not entirely dissimilar to how human coders are expected to work, and replicability of a study is here conditional on the concepts of interest being well defined and consistent, not on the textual expressions used to refer to them being identical from text to text. However, the process whereby the expressions in the collected data are matched to the concepts must remain transparent, explainable, and replicable.

For the purposes of transparency, a model must allow the analyst to inspect the features and criteria it works with during training. This is not always the case with the more recent neurally inspired models which achieve impressive but opaque performance. Recent research has addressed the task of making such models *explainable* and we can expect great advances in this area as neural,y inspired models are deployed in arenas outside research labs where accountability and transparency are important.

In addition, for application to research tasks, it is desirable that the model be *editable* to allow the researcher to manually improve its precision and recall over the data set under consideration for cases where the automatic mechanisms may have not found some correspondence of interest or where they may have overtrained on some singularity in the data set. This is in contrast with how end-to-end models typically are understood, but if the aim of a study is not to automatise concept learning but to produce an outcome to be trusted, professional human intervention is an important part of the workflow.

5. Terminological variation, referentiality, and burstiness

Topical variation reflects one of the more important aspects of language use: that of *referentiality*, where language is used to refer to items, concepts, notions of interest to discourse participants. What referential items are under consideration in a study can be predetermined through the study at hand, but may also profitably be found through a first analysis of the texts and their content: frequently, meaningful topics can emerge in textual material, unexpected to the study design, and to allow such targets of analysis to emerge is one of the primary reasons to engage in text analysis rather than purely quantitative study.

One of the most valuable features of human language is that it allows for terminological variation dependent on context. This becomes a challenge for analysis methods that rely on observations of term occurrences. Most concepts or notions of interest can be referred to with a variety of terms, and most terms may, depending on context and other variables, refer to a multitude of concepts. To ensure recall, or coverage, a method to identify concepts must have some way of finding semantically related terms, if some initial terms have been given. These may be synonyms or near synonyms (*autogiro, chopper, whirlybird*) or other related terms (*airfoil, camber, translational lift*). In a situation where interview or questionnaire data are being analysed, these sorts of recall enhancing resources should be based on general language use rather than perusal of the data at hand to avoid the pitfall of *overtraining*, especially if new data are expected to be delivered from a future iteration of the study or if the results are intended to be applied to other domains.

The general intuition of topical analysis is that some terms in language appear burstily, with local peaks in distribution, indicating that some matter of interest is under treatment. Other terms appear in a wider distribution over the entire document or document collection, constituting structural material rather than topical [21,22, e.g.]. As an example, texts which contain terms *helicopter*, *rotor*, *airfield*, and *pilot* vs texts which contain the terms *cow*, *milk*, *dairy*, and *grazing* can with some ease be classified topically by identifying terms that are unexpectedly common in a document, compared to language usage in general.

Other aspects of language use are not as obviously calculable by examining simple term counts, such as those that encode *relations* between notions that are referred to, those that organize the *structure* of the discourse, and those that indicate speaker or author *attitude*, *stance*, or *mood*.

This relates directly to variables of interest for computational social sciences: in the case study at hand, e.g. *tentativeness* vs *executive capacity*; *pragmatic* vs *principled* mindset; *empathetic* vs *solipsist* outlook and other similar personally or culturally bound concepts. Observable features that indicate such variables are sprinkled throughout the textual data and cannot be pinpointed to any single utterance, and the surface items that indicate concepts of this kind tend not to be as bursty as topical and referential items. Human readers are able to distinguish many such linguistic factors from reading

text with some precision, which means that text analysis should be able to establish observable features for them.

There is no obvious and crisp definitional demarcation between referential and topical terminological linguistic variation on the one hand and more general thematic or attitudinal linguistic variation on the other: on an operational level they vary between such items that are bursty and localised in text and such that permeate the entire body of text under consideration. The notional example documents above, with terms such as helicopter, rotor, airfield, and pilot, and documents which contain the words cow, milk, dairy, and grazing, while easy to classify by topic, reveal little of their genre, attitude of author, or perspective. They could be technical manuals, children's books, legal documents, behavioural manuals to overcome anxieties or phobias, or even volumes of poetry. For such analyses, text analysis methods must rely on linguistic items of other types than topical terms. Such stylistic analysis methods do exist, but are seldom included in text analysis packages. This is where current analysis tools risk obscuring interesting variables, by being optimised to disregard such linguistic variation which does not appear to be topical.

6. Case study

The case study which motivated us to address these more general methodological questions is an ethnographic exploratory study—"Intimacy after 60. Transition into retirement"—on women's experiences and current feelings about relationships and intimacy during their transition from working life to retirement. The study investigates how mindset and attitude towards relationships with partner, family, friends, and colleagues, with respect to compromises, principles, and conflict resolution relates to how the subjects value three aspects of intimacy: physical, emotional, and intellectual. The data collection for the study has been performed in a series of interviews in several cultural areas: North America (NA), Northern Europe (NE), Asia (A), and a selection of countries (W) characterised by recent political upheavals, low social cohesion, and deficiencies with respect to rule of law from different parts of the world. The objective of this study is to generate hypotheses for a larger more comprehensive study, and the design was formed with that in mind.

Recordings of these interviews have been manually transcribed. Each interview in the study consists of a number of lines or *turns*: the interviewer asks a question or occasionally prompts the respondent to hold forth further on some topic and the respondent then reacts to the prompt. Each such question and response ranges in length from terse answers to paragraph-length multi-sentence responses. We use turns as the basic unit of analysis. Some turns are fairly low on content; others hold various amounts of topical matter as seen in Examples (1).

(1)
a. "I forgot to mention them. They're another two that are best friends, yes. NAME1, and NAME2, and I are very close. "
b. "I think he was not a well person, because he always managed to arrange it so that I would find out. Maybe, you know, a piece... "
c. "Yes I was"
d. There is something that I talk about only with one friend, although a couple of friends know about it. Yes, but it's funny you should say that–I was recently spending a weekend with friends, NAME1 and NAME2. And NAME1 and I are very close. NAME2 and I are close, but he and I are closer. And we were walking on the beach–so tranquilizing, the water–and he said to me, he said, "You know, it's the stories that they don't tell about ourselves that are the ones that really define us."
e. "..."Oh, well, this is forever, and it'll just be the two of us." I think we got married so young because we so badly wanted to be together, but in those days it wasn't really very nice to be sexual unless you were married. So... "
f. Not necessarily being taken care of–although my friends always want to be very careful with me and kind–but I think what I want from the relationship is to not be alone, whether it's intellectual, you know, and we spend an hour bashing the president as we did the other

night, or, you know, a common interest. I have friends that I go birding with, you know? I think it's just important not to feel all by myself, not to feel too abandoned.

The data set currently under consideration consists of some 54 interviews. Overview descriptive statistics are given in Table 1.

Table 1. Descriptive statistics for the collected data.

Country	Interviews	Words	Turns	Turn length (in words)	Interview length (in words)
NA	20	88 208	3 631	24.3	181.6
A	5	21 109	1 006	21.0	201.2
W	14	2 746	1 282	18.6	91.6
NE	15	61 456	2 345	26.2	156.3

6.1. Analysis tool

The tool used in this present case study—Gavagai Explorer—is built primarily for the analysis of market surveys but which also has been used in previous academic research, [23, e.g.]. Gavagai Explorer is built to be transparent and editable and not to rely on predetermined categories or pretrained classifiers but has an extensive background knowledge of general language usage based on large amounts of previously processed text [24,25]. This enables analysts to process e.g. customer feedback, consumer reviews, or market surveys without resorting to human coders. Gavagai Explorer splits each text item—in this case, an interview turn—into sentences and clusters those sentences by terms that occur in them. Each sentence can only be in one cluster, but each turn, since it may contain several sentences, can be in several clusters. The clusters and their defining terms can be inspected and edited by the analyst, discarding spurious clusters, adding or deleting terms from existing clusters, or merging and grouping clusters into coherent sets. The tool used in this case study allows the definition of term sets to score the topical items and the topical clusters by sentiment or other attitudinal or cross-cutting conceptual themes. For each of these operations, the tool suggests synonyms and related terms to increase the coverage of the clustering stage. These synonym suggestions are obtained through a back-end lexical resource which learns semantic relations between words and multi-word terms continuously through exposure to streaming text from social and editorial media.

Each measure, whether topical or attitudinal, whether formulated through editorially determined hypotheses or discovered through analysis of occurrence patterns in the data is represented as a set of terms and scored by frequency of occurrence. We tabulate the presence of such terms in the interview turns to obtain quantitative scores for each turn and interview.

6.2. Variables of interest

For this study, the analysis has relied on manual analysis and on knowledge-based methods which are transparent as to their process model and thus are replicable and explainable; the machine learning mechanisms used by us in this study are restricted to transfer learning of lexical similarity. We will below go through the various levels of analysis and give examples for where text analysis technology is applied to the data and where we expect to be able to make use of advanced models in the near future.

1. Many variables are obviously given directly by the study design and respondent selection, through various metadata (e.g. demographic and socioeconomic variables).
2. Some variables are given as direct answers to questions posed by the interviewer. With direct answers of the type shown in Examples (2)-(4), there is a clear and codable datapoint to be found in the response. In many cases, the topic is only mentioned in the question such as in Example (3): responses can be terse or abbreviated, and refer back to the question through implicit continuation along a topic introduced by the question. This means that while the question does

not reflect the language of the respondent, text from the entire question-response pair—and in some cases, a longer stretch of discourse—holds information that the response on its own does not and which is necessary for the analysis.

These sort of responses are trivially easy for human coders to extract from the material, but are still fairly tricky for automatic analysis. In the present study, the analysis has been done by hand. Automatically processing these kinds of variables is possible with high precision for many of the cases, but extracting them automatically will entail loss of coverage, using a combination of information extraction and recent advances in sequence tagging with machine learning models. Today, the general case is not yet resolved, and the special cases that can be solved will need large amounts of training data to attain any level of reasonable coverage. Addressing this challenge promises to be a very fruitful avenue of near future research.

(2) a. INTERVIEWER: Did you work full-time?
b. RESPONDENT: I always worked full-time except for when I was in school, yeah.

(3) a. INTERVIEWER: Do you belong to any religious community?
b. RESPONDENT: No.

(4) a. INTERVIEWER: On a scale from 1 to 10, how much did your parents encourage you to get an education?
b. RESPONDENT: Not so very much. I'm going to say 4.

3. Coherence in an answer may be realised using e.g. pronominal reference to a previously named entity person. Example (5) shows how the response never mentions "husband" and thus the analyst needs to refer to the question to resolve who is posing the budgetary requests under discussion. Similarly, reference by person names is frequent in this material, as in Example (6), where the respondent uses the first name of a previous husband to refer to him.

(5) a. INTERVIEWER: Was it the same with your first husband? Was he supportive of you working full-time when you had kids?
b. RESPONDENT: Well, when I had—when we moved to PLACE in 1989, we didn't have a lot of money, and I would say that he wanted me to do more work, because I wasn't bringing in money.

(6) a. INTERVIEWER: And when you were 23 you lived for how long with ...
b. RESPONDENT: With NAME? Er, altogether maybe two or three years

Pronoun resolution—figuring out who "he" or "him" refers to locally in discourse is also a known and on a theoretical level solved task. In this case, we cannot trust such algorithms to resolve the "he" to the right referent (are we now discussing the partner, a son-in-law, a co-worker?) Similarly, identifying person names is in theory a similarly simple challenge: named entity resolution in general case is a solved task in language technology. However, in this case, resolving who NAME of the various candidate persons mentioned during the interview is involves some knowledge of the limits of the discourse at hand. Recently introduced sequence tagging models promise even better coverage than previous knowledge-based models and addressing this challenge promises, as does the previous one, to be a very fruitful avenue of near future research.

4. The variables of greatest interest for this present discussion are those expressed by the respondents in free form in unconstrained discourse. Tabulating these can be done through analysis of respondent turns in the text, and the freedom they afford the respondent are the reason to move to open responses in the first place. Much of this is fairly simple to detect. If someone mentions their partner, their husband, their boyfriend, their hubby, we know they are talking about their partner. What mentions we wish to look for and note is in a study of this sort largely governed by the hypotheses which motivated the study in the first place. A quantitative analysis of how those mentions are distributed over the material will elucidate the

relative weight of mentions of concepts as well as the correlation between such mentions: this is typically one of the primary goals of a questionnaire study in the social sciences and elsewhere. Mentioned concepts are unproblematic to identify by searching for known lexical items of interest and thus computing occurrence statistics and collocational correlations can be done with some confidence using simple string processing mechanisms. Clustering those mentioned concepts into meaningful topical bins is more of a challenge and involves some finesse in selecting clustering criteria, weighting their occurrences, and managing conceptual overlap.

For topical material, any term or set of terms with a bursty distribution is a likely candidate to be an informative topical cluster. Gavagai Explorer uses a text clustering mechanism to allow the analyst to interact with the texts in the data to cluster them into meaningful sets on which to perform quantitative analyses, merging similar clusters, discarding spurious clusters, and directly editing the set of terms that define the cluster. Samples of two automatically grouped topics are shown in Examples (7) (clustered around various leisure activities) and (8) (clustered around the concept of friendship). It is a non-trivial challenge for human analysts to provide an exhaustive list of variants such as "hubby", "beau", etc for "partner" and here technology can help by suggesting synonyms. In this study we have used a learning back-end synonym lexicon, trained from a continuous stream of text material, to suggest lexical items of relevance for this purpose.

(7)
a. I love theater and movies.
b. I travel with my daughter, my cousin, and the last trip I did I did alone.
c. Movies I also like and I do knit.
d. Netflix is wonderful, we have a glass of wine and peanuts and a movie.
e. He did lots of traveling.
f. I do yoga two times a week and I dance three times a week and I do gymnastics.
g. We do concerts and movies together.

(8)
a. With my friends I had more freedom.
b. Lately, I let go of friends that don't work out anymore.
c. Different groups where I have my best friends.
d. I talk about my issues with my best friends.
e. Now he has a girlfriend from PLACE
f. I am intellectually close to my friends and also emotionally.
g. I really enjoy talking to my friends about their experience of politics, my friend used to live in PLACE, he lived there for years during Pinochet.

5. Some variables cut across topic, such as the sentiment shown by the respondent to a topic which is mentioned in the interview. A topic can be mentioned in a positive or negative tone, with skepticism, revulsion, anger, or frustration. The palette of human emotion is manifold (and its composition and parameters of variation are under continuous discussion [26–33]); in a study such as the present example, some expressions of sentiment are pertinent to the research hypotheses. Example (9) gives some samples of items with negative sentiment from the interviews.

(9)
a. I'm so very sorry. I didn't expect to be talking about all this horrible stuff. Thank you.
b. I'm sad hearing that. The first one I think was okay, but the second two... Because then you're using sex...
c. Of course they treated her terrible, right?
d. I believe that he married her to have access to the children, because he was a monster.
e. So I became miserable there.
f. But he verbally abused me horribly.
g. It was terrible, it was terrible, and nobody knew.

Sentiment analysis is a known technique in language technology, and while sentiment analysis is less mature than many of the techniques, it is eminently useful in this case, where the items under consideration—the turns—are clearly delimited in scope. In this study, we have made use of a commercial multivariate sentiment analysis functionality which compares well in head-to-head comparisons to other existing models. This is an area which is in rapid development, and we

found the possibility to define expressions of feelings and sentiments to suit the hypotheses of the study such as "Feeling rejected" or "Taking initiative" to be valuable addition to standard positivity and negativity.

6. Finally, the data set allows us to assess variables that permeate an entire set of responses from some respondent. Some of the hypotheses of the current study call for e.g. distinguishing pragmatic respondents who put effort into compromise and conflict resolution from respondents who hold their social circles to set standards and whose social actions are bound by explicitly formulated principles; respondents with a positive outlook from those who are more gloomy; respondents who express emotions with intensity from those who are more reticent; and other similar personally or culturally bound concepts. In contrast with sentiment, which is here understood as an attitude shown vis-à-vis some mentioned topic, mindset markers are sprinkled throughout responses, cannot be found in any single turn, and need to be aggregated over an entire session.

These sorts of variables are to a some extent possible to compute from lexical statistics alone. If a respondent repeatedly uses terms to indicate bitterness, uncertainty, or exuberance, these can be aggregated to provide measures of interest. In this study we have used the sentiment scores discussed above, and tabulated the difference between number of observed expressions for various positive sentiments and for various negative sentiments as a measure of *polarity* of emotion for individuals and a sum of squares for both as a measure of *intensity* of emotion for individuals.

6.3. Results

Initial clustering shows a number of statistically solid clusters of potential interest. These are examined and then iteratively refined by the above operations: merging, discarding, reformulating. For most studies, the topics of interest are largely informed by hypotheses and of course strongly determined by what topics are brought up in the interview and this one is no exception. The topics found in the text, after alignment with project hypotheses, have to do with various types of intimacy and closeness: physical, intellectual, emotional, sexual, and various interpersonal relations of the respondent: partner, friends, colleagues, family, children, pets.

In the present case study, measures such as rejection, situational and pragmatic stance, adherence to principles, taking initiative, getting along, conflict resolution, and feeling important were defined to capture personality traits and social stance of the respondent.

Using these measures, we can then assess the relative importance of the various types of closeness, the differential between the various cultural areas, and the individual variation. We can measure the amount of attention spent on the various interpersonal relations and the attitude towards them. We can see if respondents from some cultural area tend to score differently than others on some of those measure, and if personality type or mood, assessed by general tendency to score higher or lower on attitudinal measures correlates with attitude towards some topic of interest. Some sample results are given here.

Table 2. Difference in emphasis on the various aspects of intimacy across cultural areas (percentage of all turns that bring up the facet of intimacy in question)

Cultural area	Physical	Emotional	Intellectual	Sexual
all	6.34%	6.64%	4.85%	2.61%
NA	8.32%	5.98%	5.34%	3.30%
A	6.86%	6.46%	5.57%	1.49%
W	9.44%	8.11%	6.90%	2.50%
NE	1.36%	6.95%	2.69%	2.09%

In Table 2 we show the varying emphasis on the aspects of intimacy studies across cultural areas. We count what proportion of the respondent turns mention the aspect of intimacy in question. These can the further be analysed to examine differences in attitude: which areas are more positive or negative with respect to the intimacy aspects in question. We find here that the cultural area has bearing on these factors: respondents from Northern Europe place much less emphasis on physical and intellectual intimacy than other respondents do. Similarly, as shown in Table 3 we find that hedged or cautious expressions with overt markers of skepticism are more prevalent and intensity of expression is less pronounced in responses from Northern European respondents than in others and that cultural areas where political upheaval has been present show the opposite pattern.

Table 3. Cultural areas and attitude in text (scores computed based on weighted occurrence of polar terms, attitudinal terms, and hedge terms)

Country	Polarity	Intensity	Skepticism
NA	26.8	193.0	11.1
A	21.6	163.3	11.0
W	11.8	256.9	2.64
NE	10.1	66.3	22.8

Across the entire data set, for all individuals, we find e.g. that placing high importance on physical intimacy is well correlated with placing high importance on intellectual intimacy, but that importance of emotional intimacy correlates less with the other facets of closeness. We also find that those who express themselves with more positive than negative terms more often bring up the concept "Getting along" than others. Similar correlations can be investigated over all items under study.

7. Conclusions, Lessons learnt, and Paths Forward

The computational social science case study, which is currently under continued execution and analysis, uses text analysis technology for processing responses from questionnaires. The tool used is developed for the analysis of customer feedback and related data which task has obvious similarities to processing data for the social sciences. We found that topical clustering and terminological enrichment provides for convenient exploration of the responses: tabulating observed concepts allows for correlation analysis and inference of relations between demographic data and attitude towards concepts of interest for the study. This makes rapid hypothesis testing and comparison between textual and non-textual variables possible. Studies in social science have great potential to allow for more exploratory open-ended studies with less effort, increase coding consistency, and reduce turnaround time for the analysis of collected data by using tools developed for market purposes.

The tools, such as the one used in the present study, that are available for text analysis today are not tailored to the task of questionnaire material, however. While many of the variables under consideration are quantifiable using lexical statistics, we find that some interesting and potentially valuable features are difficult or impossible to automatise reliably at present. We especially note as potentially quite useful avenues of further investigation that:

1. Topical information is relatively straightforward to extract using lexical statistics whereas attitudinal content in many cases poses greater challenges. There is a large and growing body of work on attitude and sentiment analysis in text analysis, but in most cases, the palette of emotions or attitudes will need to be tailored to the needs of the study at hand, which will require more work on the part of the analyst.
2. Features that permeate the entire text of an interview are more challenging to extract. For instance, in this case study, establishing the mindset of the respondent with respect to their social circles, whether they are principled and rule-bound or whether they tend towards compromise and pragmatism. Extracting this information from the interview text is not difficult for a human coder, but defies lexically based computational efforts. Automating such extraction using recent

machine learning models is again quite possible, but needs specific targeted efforts tailored for this purpose.

3. The interplay between questions and responses poses specific requirements for text analysis which are possible to address using today's technology if correspondences of interest are used to train a model.This as to our knowledge not yet been attempted.
4. Traditional natural language processing mechanisms developed for general application in text analysis such as named entity recognition and anaphor resolution are excellent candidates to improve recall for questionnaire processing if their range and scope are tailored to the specifics of question-response interplay.

Most crucially, any analysis model used for computational scholarship tasks where previously human effort has been the major analysis mechanism, must be *transparent*, its deliberations must be *modifiable*, and its decisions must be *explainable*. These requirements are necessary both to afford the researcher trust in the analysis results and for review, reuse, and replication by others.

We conclude that using text analysis tools to process material for computational social science is a most definitely useful path of further investigation, and that text mining and related technologies have their place in knowledge discovery here as in other fields of study. Situations where researchers in social sciences hesitate to include open answers in questionnaires or are wary of processing large amounts of textual data can well be met using text analysis technology, with today's tools and even more so with coming tools.

Author Contributions: Conceptualization, all; methodology, all; software, Karlgren; validation, Meyersson Milgrom; formal analysis, Karlgren and Li; investigation, Meyersson Milgrom; data curation, Li; writing—original draft preparation, Karlgren; writing—review and editing, all; project administration, Meyersson Milgrom; funding acquisition, Karlgren and Meyersson Milgrom

Funding: Karlgren's work was done as a visiting scholar at the Department of Linguistics at Stanford University, supported by a VINNMER Marie Curie grant from VINNOVA, the Swedish Governmental Agency for Innovation Systems (2016-040601). Meyersson Milgrom's work has been supported by the Trione director's Fund, Stanford Institue of Economic Policy Research, Stanford University

Conflicts of Interest: Karlgren is at time of writing employed at Gavagai, a text analysis company which builds and sells one of the tools used in the study.

References

1. Singer, E.; Couper, M.P. Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)* **2017**, *11*, 115–134.
2. Blair, E.; Sudman, S.; Bradburn, N.M.; Stocking, C. How to ask questions about drinking and sex: Response effects in measuring consumer behavior. *Journal of marketing Research* **1977**, *14*, 316–321.
3. Lazarsfeld, P.F. The art of asking WHY in marketing research: three principles underlying the formulation of questionnaires. *National marketing review* **1935**, pp. 26–38.
4. Dohrenwend, B.S. Some effects of open and closed questions on respondents' answers. *Human Organization* **1965**, *24*, 175–184.
5. Schuman, H.; Presser, S. The open and closed question. *American sociological review* **1979**, pp. 692–712.
6. Wenemark, M.; Hollman Frisman, G.; Svensson, T.; Kristenson, M. Respondent satisfaction and respondent burden among differently motivated participants in a health-related survey. *Field Methods* **2010**, *22*, 378–390.
7. O'Cathain, A.; Thomas, K.J. " Any other comments?" Open questions on questionnaires–a bane or a bonus to research? *BMC medical research methodology* **2004**, *4*, 25.
8. Macskassy, S.A.; Banerjee, A.; Davison, B.D.; Hirsh, H. Human Performance on Clustering Web Pages: A Preliminary Study. KDD, 1998, pp. 264–268.
9. Adams, S. The Office of Science Information Services, National Science Foundation. *Bulletin of the Medical Library Association* **1959**, *47*, 387.
10. Sager, N. Syntactic analysis of natural language. In *Advances in computers*; Elsevier, 1967; Vol. 8, pp. 153–188.

11. Grishman, R.; Sundheim, B. Message understanding conference-6: A brief history. The 16th International Conference on Computational Linguistics (COLING). International Committee on Computational Linguistics, 1996.
12. Grishman, R. Information extraction: Techniques and challenges. International Summer School on Information Extraction. Springer, 1997.
13. Cowie, J.; Wilks, Y. Information extraction. *Handbook of Natural Language Processing* **2000**, *56*, 57.
14. Aggarwal, C.C.; Zhai, C. *Mining text data*; Springer Science & Business Media, 2012.
15. Blei, D.M.; Lafferty, J.D. Topic models. In *Text Mining*; Chapman and Hall/CRC, 2009; pp. 101–124.
16. Fitzpatrick, K. The Humanities, Done Digitally. *The Chronicle of Higher Education* **2011**.
17. Moretti, F. *Distant reading*; Verso Books, 2013.
18. Da, N.Z. The Digital Humanities Debacle—Computational methods repeatedly come up short. *The Chronicle of Higher Education* **2019**.
19. Underwood, T. Dear Humanists: Fear Not the Digital Revolution. *The Chronicle of Higher Education* **2019**.
20. Da, N.Z. The Computational Case against Computational Literary Studies. *Critical Inquiry* **2019**, *45*.
21. Katz, S.M. Distribution of content words and phrases in text and language modelling. *Natural language engineering* **1996**, *2*, 15–59.
22. Madsen, R.E.; Kauchak, D.; Elkan, C. Modeling word burstiness using the Dirichlet distribution. Proceedings of the 22nd international conference on Machine learning. ACM, 2005, pp. 545–552.
23. Georg, C. Virtual patients in nursing education: teaching, learning and assessing clinical reasoning skills. PhD thesis, Karolinska Institutet, Stockholm, 2019.
24. Sahlgren, M.; Gyllensten, A.C.; Espinoza, F.; Hamfors, O.; Karlgren, J.; Olsson, F.; Persson, P.; Viswanathan, A.; Holst, A. The Gavagai living lexicon. Language Resources and Evaluation Conference. ELRA, 2016.
25. Espinoza, F.; Hamfors, O.; Karlgren, J.; Olsson, F.; Persson, P.; Hamberg, L.; Sahlgren, M. Analysis of Open Answers to Survey Questions through Interactive Clustering and Theme Extraction. Proceedings of the 2018 Conference on Human Information Interaction&Retrieval (CHIIR). ACM, 2018, pp. 317–320.
26. Darwin, C. *The Expression of the Emotions in Man and Animals*; John Murray: London, 1872.
27. James, W. What is an emotion? *Mind* **1884**, pp. 188–205.
28. Mehrabian, A.; Russell, J.A. *An approach to environmental psychology*; M.I.T. Press: Cambridge, Massachusetts, 1974.
29. Morgan, R.L.; Heise, D. Structure of Emotions. *Social Psychology Quarterly* **1988**, *51*, 19–31.
30. Ekman, P. An argument for basic emotions. *Cognition and Emotion* **1992**, pp. 169–200.
31. Kuppens, P.; van Mechelen, I.; Smits, D.J.M.; de Boeck, P. Associations Between Emotions: Correspondence Across Different Types of Data and Componential Basis. *European Journal of Personality* **2004**, *18*, 159–176.
32. Coan, J.A.; Allen, J.J.B., Eds. *The Handbook of Emotion Elicitation and Assessment*; Oxford University Press, 2007.
33. Izard, C.E.; King, K.A. Differential emotions theory. In *Oxford Companion to the Affective Sciences*; Scherer, K., Ed.; Oxford University Press, 2009.