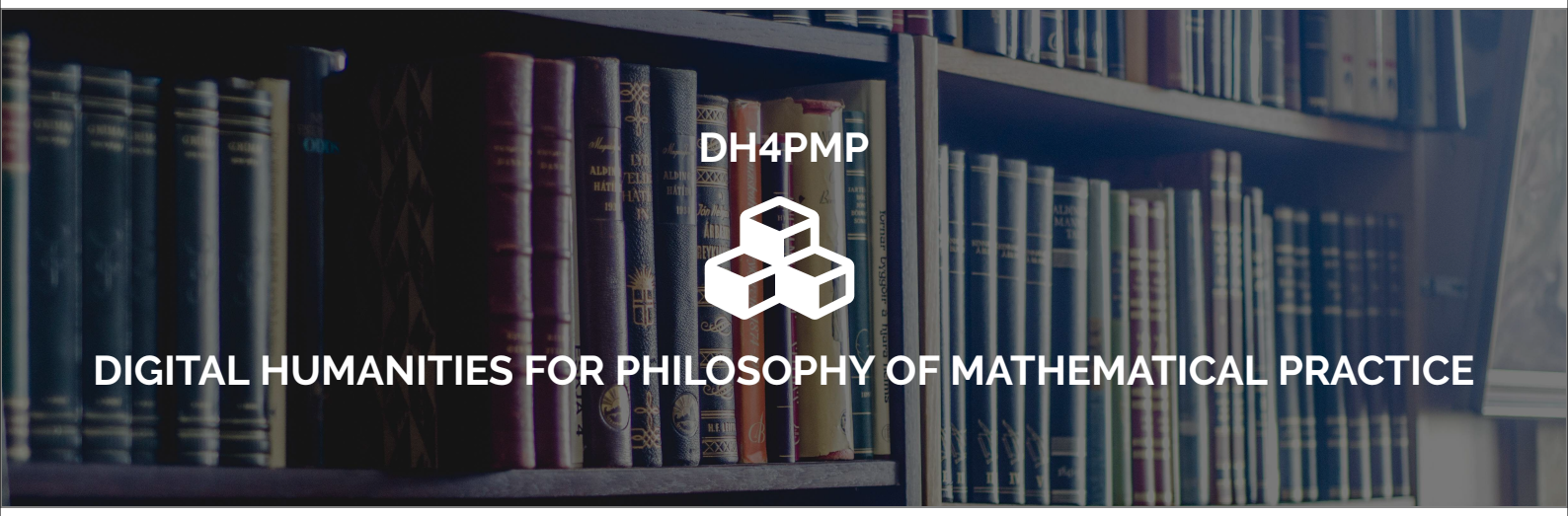

Binary classification of MR abstracts

Henrik Kragh Sørensen

Digital Humanities for Philosophy of Mathematical Practice, Section for History and Philosophy of Science, Department of Science Education, University of Copenhagen, Denmark. henrik.kragh@ind.ku.dk.

May 3, 2021

Acknowledgements

I am grateful to Sophie Kjeldbjerg Mathiasen for the collaboration and discussions that shaped this note.

Status & caveats on this note

Version 1 includes preliminary data on the training of transformer classifiers; I am continuing to run training and evaluation.

1 Introduction

The contextual study of new corpora holds great promise for developing new insights into mathematical practice through the use of digital humanities. As made famous by Franco Moretti and first employed in literary studies, the notion of *distant reading* underpins much of the methodological promise of digital humanities. In direct opposition to *close readings*, a distant reading focuses on analysing texts *without* going deeply into close contextual interpretations. But whereas the first wave of digital humanities relied on rather coarse and 'blind' statistical analyses, the advent of powerful tools in artificial intelligence in general and natural language processing in particular, are allowing researchers to gather meaningful analyses from large corpora. To incorporate methods of digital humanities into the already methodologically heterogenous, emerging subfield of philosophy of mathematical practice, we are looking to *gain from* distant reading in *answering philosophical questions* (see Figure 1).

Among the most fascinating corpora to now lend themselves to new kinds of philosophical interests are the institutionalized abstracting services of the *Mathematical Reviews (MathSciNet, formerly MR)* and *Zentralblatt Mathematik (zbMath)*. These corpora offer a new fruitful and

Henrik Kragh Sørensen (Mar. 12, 2021). *Binary classification of MR abstracts.* DH4PMP Work-in-Progress & Proof-of-Concept Series 9.

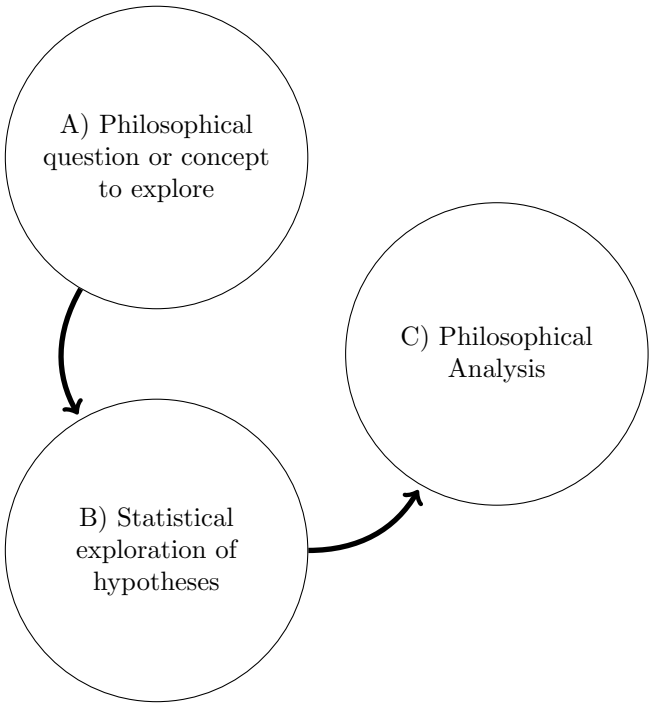

Figure 1: The triangular methodological layout, starting from a philosophical question, addressing it using methods of digital humanities, to draw philosophical analyses.

largely untapped source for enquiry into what we may call *secondary* mathematical communication.[1] Whereas *primary* mathematical communication is embodied in research papers and monographs, the reviews offer a *secondary* perspective as a mathematician — different from the original author — is offering a short review or abstract of the original work, including evaluations and comments.

To tap this source, we could pursue an empirical research design as outlined in Figure 2. To access the `MathSciNet`, we first define a limiting query that defines a corpus, defined by our research question. We can then sample from the corpus for various purposes, including exploratory, qualitative, or quantiative studies. For instance, we may use exploratory and grounded qualitative methods to form categories to pursue for philosophical analysis.[2]

However, since the corpora are prohibitingly large for comprehensive analysis, it can be valuable to use digital tools in filtering the sample or help highlight key features for humans to review. And luckily, recent tools in natural language processing, including text classification in general and the advent of *transformers technology* in particular, can be of assistance.

In this note, I report on efforts to assist human classification by using text classifiers to eliminate overly dominant and recurring features from the grounded analysis (the green pentagon in Figure 2). Subsequently, in order to hypothetically assess the extent of the class in the 'great unread' part of the corpus, I trained potentially more powerful classifiers, and a report on their results and an estimation of the global prevalence of so-called *numerical verification* in reviews

[1] For more discussion of how to mine these corpora, see **dh4pmp:wippoc:10**.

[2] It is then also possible to attempt to instrumentalize the categories and quantitatively test hyposes (the red part of Figure 2), but this is of less relevance here.

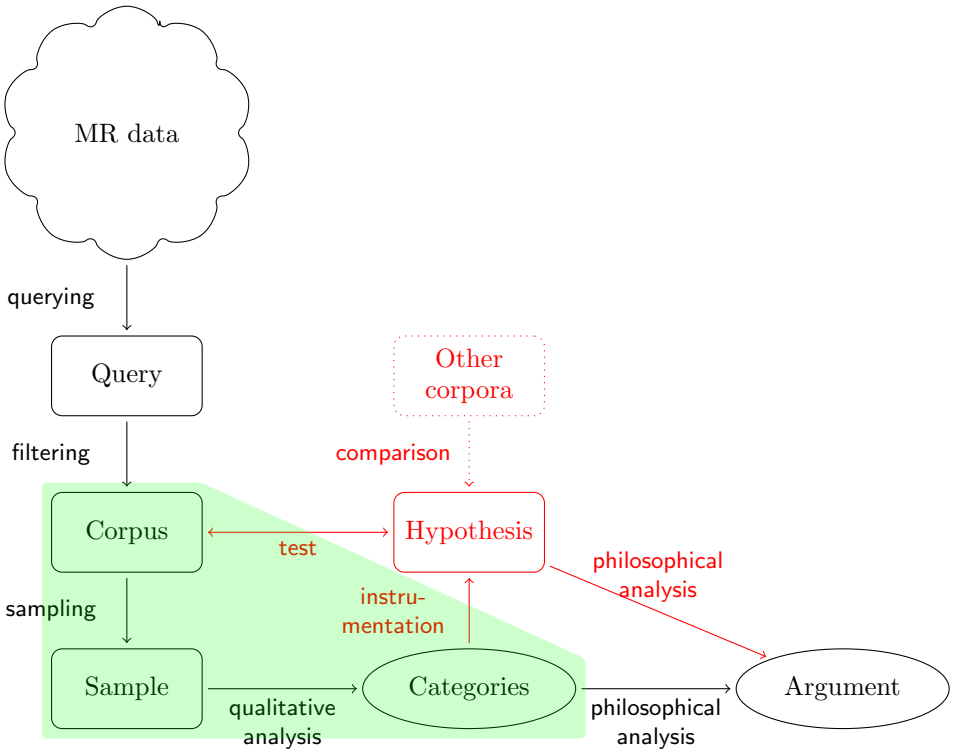

Figure 2: Schematic overview of empirical research design in corpus analysis.

concerning mathematical experiments is given.

2 Method

2.1 Sampling and labelling

As part of her Master's thesis, Mathiasen (2021) undertook a qualitative, grounded classification of reviews from `MathSciNet` in which the work "experiment" (or its derivative) occurred. Her work was based on a corpus derived from `MathSciNet` containing a total of 94,732 reviews (as of November 2020) meeting the search criterion.

For each review in the corpus, metadata is automatically joined from the search of the `MathSciNet`. In particular, the variable `mscmain` was adjoined to each review, containing the primary, main Mathematical Subject Category associated with the work under review.[3] Furthermore, when the actual review is required, the LaTeX of it is massaged so as to retain only running text.[4] When reviews are exported for hand-coding, key words, in paritcular derivations of 'experiment' and 'numerical' could be highlighed.

In a first, exploratory run, Mathiasen looked at samples drawn from a small number of MSC categories of which she was unsure whether they fitted into her analyses. This set is now referred to as Pilot #1, and it consisted of a total of 5×50 reviews from the MSC categories 68, 90, 91, 93, and 94.

Among these reviews, Mathiasen (2021) observed that a large number dealt with experiments in the sense of 'numerical verification', and it seemed that this category (referred to now as NV)

[3]For instance, for a work categorized as 51P05 (05D05 82B05), 51P05 would be its primary (first) category, and 51 would be the main part of this, thus `mscmain` for this work would be 51.

[4]We now have stronger tools for this, but here, it can be done almost exclusively using regular expressions.

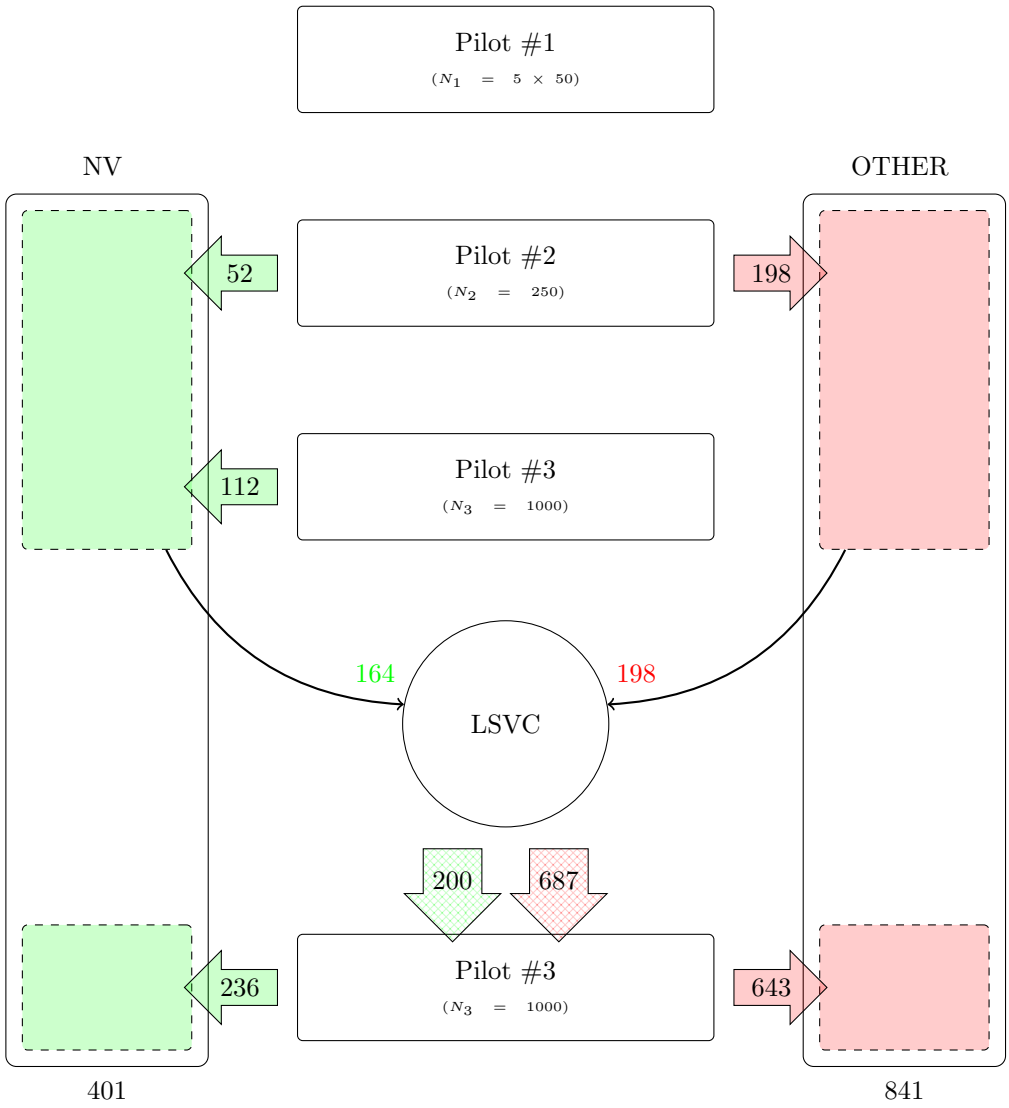

Figure 3: Building the training set.

ran the risk of watering down all other interesting categories for her study. As a result of her considerations, the she chose to focus on six main MSC categories for her subsequent analyses.

To begin the grounded analysis of types of use of the word "experiment" in mathematics, a second sample (Pilot #2) was drawn, consisiting of 250 reviews from the set of positively included MSC categories from the following set,

$$\{35, 65, 62, 68, 90, 93\}.$$

Each of these reviews was coded by hand, resulting eventually in five main types and an auxiliary type. Among the five types is also NV, and in Pilot #2, Mathiasen identified 52 reviews belonging to NV. The remaining 198 were all confirmed as *not* belonging to NV, thus establishing a partition of Pilot #2 into two labels, NV and OTHER.

2.2 Inference with LSVC

At this point, we (the present author and Mathiasen) decided to try to use machine-learning techniques to filter out NV from the next pilot. To this end, once we had drawn the sample for