Lancaster University Management School

2nd Workshop on Textual Analysis Methods in Accounting and Finance

Lancaster University Management School, 3-5 September 2019

Provisional programme

Day 1: 3 September

11.00-11.15 Welcome and introduction

11.15-12.30 **Session 1** ***Overview of textual analysis literature in accounting and finance***
The aim of this session is to provide participants with an overview of extant research on textual analysis in the accounting and finance literature. We will focus on the proposed benefits of automated analysis of text and evaluate extant research against these perceived advantages. A key conclusion that will emerge from the review is that prior research is limited in scope and fails to deliver many of the suggested benefits. A critical theme informing the remainder of the workshop is that automated analysis is not a "quick fix" replacement for close manual reading by domain experts: most advanced applications of computational methods rely on significant manual reading for training and validation.

12.30-13.15 <Lunch>

13.15-15.15 **Session 2** ***Text extraction and preprocessing: The basics***
Automated text retrieval is the starting point for most large-sample applications of textual analysis in accounting and finance. This session will provide general guidelines on the text retrieval process and hands-on experience with retrieving: 10-K annual report text (including harvesting documents from EDGAR) using python and R scripts; UK annual report narratives published as PDF files using the CFIE-FRSE annual report app; and earnings announcement narratives using the CFIE PEA app. The session will also review the text preprocessing choices that researchers should consider prior to computing text features.

15.15-15.30 <Break>

15.30-17.30 **Session 3** ***Readability and tone: Methods and critique***
Readability and tone (sentiment) are the two most commonly analysed features of financial market text. This session will review and critique methods used in the extant literature to measure readability and tone. We will demonstrate the problems of relying on standard readability metrics such as Fog to capture sophisticated narrative features such as complexity and understandability. We will also review the various approaches for measuring tone, ranging from simple wordlists to more advanced machine learning methods. A key conclusion that will emerge from this review is that simple measures of readability and tone provide limited scope for generating significant new insights in the literature.

18.00-19.30 **Dinner & research presentation:** ***Classifying Tone and Attribution***
A buffet dinner followed by a discussion of ongoing research assessing the relative accuracy of wordlists and machine learning for measuring tone and managerial self-attribution bias in performance sentences from earnings announcements.

Day 2: 4 September

09.00-10.30 **Session 4 *Constructing and using wordlists***
Wordlists are the most common approach to analysing financial text in the accounting and finance literature. This session discusses the advantages and weaknesses of using a wordlist approach to study financial text, reviews the most common wordlists employed in the literature, and considers some of the methods used in conjunction with wordlists to improve their classification performance. The session will also explain the different approaches to constructing wordlists.

10.30-11.00 <Break>

11.00-12.30 **Session 5 *Introduction to machine learning***
While machine learning forms the basis for a large proportion of research in the field of natural language processing, its uptake in accounting and finance is more limited. This session provides a board introduction to the field of machine learning methods, including both supervised and unsupervised approaches. Different aspects of machine learning and their relation will be explained including classification, named entity recognition, summarization, and topic modelling.

12.30-13.30 <Lunch>

13.30-15.00 **Session 6 *Machine learning applications: Classification***
This session provides a hands-on introduction to classification using machine learning methods. Participants will use the Weka toolkit (https://www.cs.waikato.ac.nz/~ml/weka/downloading.html) to construct and evaluate a model for identifying fraudulent financial reporting using 10-K filings. Results and insights from the analysis will be used to highlight weaknesses in the extant literature and identify opportunities for future research.

15.00-15.15 <Break>

15.15-17.15 **Session 7 *Machine learning applications: Topic modelling***
Several recent papers in the accounting literature have employed topic modelling methods such as Latent Dirichlet Allocation (LDA) to identify topics in financial text. This session provides a hands-on introduction to topic modelling. Participants will use MALLET (http://mallet.cs.umass.edu/index.php) to extract topics from an annual report corpus. In addition to walking participants through the pracitcalites of the modelling process, the session will highlight the many problems associated with topic modelling and discuss alternative approaches to the content analysis problem.

18.00-19.30 **Dinner & research presentation: *Characteristics of Award Winning Annual Reports***
A buffet dinner followed by a discussion of ongoing research that employs corpus methods to identify topics and linguistic styles that characterize high quality annual report narratives (proxied by reports shortlisted for a reporting award).

Day 3: 5 September

09.00-10.30 **Session 8** ***Introduction to corpus linguistics***
This session provides an introduction to the theory and core methods underpinning the systematic analysis of a large body of text (i.e., a corpus). The session will cover the following themes: introduction to basic corpus linguistic concepts; presentation of different corpora types and examples; methodology for corpus design, compilation, and processing; corpus annotation; basic resources and corpus analysis tools; examples from the literature of using corpus methods to analyse analysis of financial discourse.

10.30-11.00 <Break>

11.00-13.00 **Session 9** ***Applied corpus methods: Tools and techniques***
This session provides hands-on experience of corpus analysis. The session will consist of two parts. Part 1 will introduce the corpus that will form the basis of our analysis (Brexit narratives in annual reports of UK financial firms), along with the #LancsBox software for corpus analysis. In Part 2, participants will use #LancsBox to analyse a small dataset and perform corpus tasks including: extracting word lists; finding collocates; and searching for n-grams and keywords. The session will conclude with a discussion of the insights gained from analysing the corpus.

13.00-14.00 <Lunch> and workshop ends

14.00-15.30 Optional surgery session for PhD students seeking feedback on research proposals and ongoing work involving analysis of text