Applying machine learning in accounting research

Machteld Van den Bogaerd *, Walter Aerts

Department of Accounting and Finance, University of Antwerp, Faculty of Applied Economics, Prinsstraat 13, B-2000 Antwerp, Belgium

ABSTRACT

Quite often, in order to derive meaningful insights, accounting researchers have to analyze large bodies of text. Usually, this is done manually by several human coders, which makes the process time consuming, expensive, and often neither replicable nor accurate. In an attempt to mitigate these problems, we perform a feasibility study investigating the applicability of computer-aided content analysis techniques onto the domain of accounting research.

Krippendorff (1980) defines an algorithm's reliability as its stability, reproducibility and accuracy. Since in computer-aided text classification, which is inherently objective and repeatable, the first two requirements, stability and reproducibility, are not an issue, this paper focuses exclusively on the third requirement, the algorithm's accuracy. It is important to note that, although inaccurate classification results are completely worthless, it is surprising to see how few research papers actually mention the accuracy of the used classification methodology.

After a survey of the available techniques, we perform an in depth analysis of the most promising one, LPU (Learning from Positive and Unlabelled), which turns out to have an *F*-value and accuracy of about 90%, which means that, given a random text, it has a 90% probability of classifying it correctly.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Quite often, in the hope of discovering new relations and insights, research involves analyzing and classifying large quantities of unstructured information, such as news items or annual reports. So, to accurately extract the relevant information from the entire, and possibly huge, set of available raw data is an important challenge. Since this can easily mean analyzing multiple thousands documents, manual analysis is often not feasible, unless off course, one limits himself to a very small subset of the total amount of available data.

Therefore, researchers typically delegate this very time consuming task to a series of human coders who will look through the entire set of news items and perform the classification. There is however a problem. You know, either implicitly or explicitly, which characteristics will make a document interesting, or positive, for you. The difficulty is conveying these criteria to your coders. But often, you will classify those documents, at least partly, on intuition. When you attempt to explain these intuitive criteria to your coders, some nuances will be lost and their results will never completely match your intents. Or, put more scientifically, the accuracy with which they classified the documents the way you intended it, will almost never be 100%. This is only achievable if you do the job yourself. Furthermore, repeatability is a major concern

as a different group of coders will almost certainly produce a different outcome.

In an attempt to decrease the cost and further increase the speed of classification, recent research delegates this highly time consuming process to computers instead of human coders. Without actually being able to argument the process, humans are able to make fussy decisions, like for example, whether a news item is favourable or unfavourable. Classification algorithms basically try to autonomously discover the implicit, underlying decision structure used by their human counterpart. As such, one can say that these algorithms try to mimic human decision making behaviour. However, this introduces an important limitation. Most computer aided classifications are limited to the use of naïve, heuristic algorithms such as word counting and frequency analysis. The problem with these techniques is that they are not very accurate as they are not capable of handling the many syntactical and semantic nuances one typically finds in natural languages. And because of this, they suffer from a rather large number of false positives and false negatives (Nigam, Mccallum, Thrun, & Mitchell, 2000).

A more modern, and better, approach is the use of machine learning techniques. In this case, a self-learning algorithm tries to discover the underlying characteristics on which a human classifier classifies documents as either relevant or not. In a second phase it utilized this newly acquired knowledge to autonomously classify the entire set of documents.

In this contribution, we present a methodology for using these advanced machine learning techniques in accounting research.

* Corresponding author. Tel.: +32 3 2654109; fax: +32 3 2654064.
  E-mail address: machteld.vandenbogaerd@ua.ac.be (M. Van den Bogaerd).

More specifically, we demonstrate that the use of a text classification algorithm will greatly improve the classification speed of unstructured information while maintaining a very high accuracy.

The paper is structured as follows: Section 2 describes the major text classification methodologies; Section 3 describes a series of text classification application, both general and specific to accounting research. Next, Section 4 presents an overview of the most important text classification approaches. Section 5 maps these approaches to two candidate algorithms and selects the most appropriate one, in this case LPU (Learning from Positive and Unlabelled) which is discussed in detail in Section 6. Next, Section 7 describes the complete process from raw data to classified information. Section 8 presents three possible algorithms for automatically building test sets. Section 9 describes the software program written to support this process and Section 10 evaluates the proposed classification process. To wrap up, in Section 11 we summarize the complete process in four practical steps and finally present a real-life application in Section 12.

2. Text classification methodologies

Neuendorf (2002) defines content analysis as the systematic, objective, quantitative analysis of message characteristics. Due to the rapid advancements in computer text content analysis software, content analysis has been one of the fastest growing techniques of the last 20 years, especially in the field of mass communication research.

Text classification can be seen as a special case of content analysis. It is a rather young discipline which tries to classify an artefact into one or more classes. Usually, for simplicity, the artefact will be a written document, but this is not a requirement. Also, most algorithms will limit the number of output classes to just one. Either the artefact is part of this class, or it is not (Nigam et al., 2000).

Text classification methods can be divided into three major categories (Rosenberg, Schnurr, & Oxman, 1990):

○ Individual word-count systems: this kind of coding is generally done by a computer because of the simplicity and repeatability of the task. All the words receive a tag or category description, which is linked to procedures like frequency counts and other statistical procedures to identify text characteristics.
○ Human-scored, phrase-based content analysis: here human coders first divide text into phrases, which are then classified based on an explicit set of classification rules. In accounting research this method is mostly used in analyzing social disclosure like, e.g. environmental disclosures, intellectual capital disclosures (Beattie & Thomson, 2007; Brammer & Pavelin, 2006; Deegan & Gordon, 1996; Gamble, Hsu, Kite, & Radtke, 1995; Hasseldine, Salama, & Toms, 2005; Patten, 2002; Tilt & Symes, 2000), etc.
○ Computerized systems or CAQDAS (Computer-Aided Qualitative Data Analysis Systems): the newest systems contain parsers or artificial intelligence features to combine both syntax and lexicon in the text analysis (Kelle, Prein, & Bird, 1995).

An important question to ask is how well these different systems perform. For this, Krippendorff (1980) identifies three types of reliability: stability, reproducibility and accuracy. The stability of an analysis is high if a coder has the same results for the same analysis twice (with a certain time-gap). Reproducibility is the extent to which the coding produces the same result when the text is coded by multiple coders. And accuracy is the extend to which the obtained result coincides with the optimal result.

Milne and Adler (1999) and Unerman (2000) investigated the reliability of the results when applying one of the first two categories and found that most research scored very badly. So, either the results were not very stable, reproducible or accurate. When using CAQDAS on the hand, as coding is automated and done on an objective and a repeatable way, these systems do not suffer from stability or reproducibility problems. However, as it is much more difficult for a computer than a human coder to understand the meaning of a text, these approaches often suffer from a low accuracy.

Literature, summarized below, shows that most researchers seem to have a preference for one of the two manual methods, which is a petty as a computerized approach has the potential of greatly increasing the reliability of the results. In an attempt to introduce CAQDAS approaches to accounting research, we will show the results of a very extensive accuracy test to show that it is possible to use automated algorithms, which do not suffer from stability or reproducibility problems, save huge amounts of time and effort, and have a very high accuracy.

3. Text classification applications

Many examples of text classification applications can be given. The best known one is spam filtering. Here, a text classification algorithm tries to determine whether a message is spam or not (Yerazunis, Chhabra, Siefkes, Assis, & Gunopulos, 2005). Another common use is while searching the internet for documents in a certain language. The search engine will use text classification algorithms to decide in which language a given document is written. In agents based search algorithms, which is a technology developed to help human observers overcome information overload, an autonomous software application will go out on the internet to search for information that could be of interest to his owner (Cawsey, 1997). In military, classification techniques are used to analyze both written and spoken conversations and to determine whether they are suspicious or not.

When zooming in on economics research, many applications of text classification can be found. In the field of direct marketing: a core problem is selecting, and investing time, in those individuals which are most likely to become future customers. A possible solution is to build a profile of your current customers and to classify the complete set of potential customers according to whether they fit this profile or not. Those who match the profile are the most likely future customers (Liu, 2006).

In the domain of corporate social responsibility (CSR) disclosures, and more specific environmental disclosures, researchers have repeatedly applied classification algorithms. Researchers like Hogner (1982), Gray, Kouhy, and Lavers (1995a, 1995b), Deegan and Rankin (1996), Brown and Deegan (1998) and Neu, Warsame, and Pedwell (1998) used classification algorithms to analyze environmental information in annual reports. While Brown and Deegan (1998) analyze media news items of all firms with environmental impact, Deegan and Rankin (1996) focus on disclosures of companies who had been prosecuted by the Environmental Protection Authority.

When looking at other accounting research, Abrahamson and Park (1994) and Abrahamson and Amir (1996) examined the content of the president's letters to shareholders by searching for the number of negative words in these letters. Deephouse (2000) introduced content analysis in the measurement of a firm's media reputation. Previously, media reputation was measured by using the theoretically weak Fortune rankings, but Deephouse changed the scene by using the 'media favourableness coefficient', which was based on an intensive rating process, where all news items about certain banks were manually classified as favourable or unfavourable. Pollock and Rindova (2003) and Greenwood, Li, Prakash, and Deephouse (2005) followed his example, by using the same method to measure reputation.