Text Mining: An introduction to theory and some applications

Nadir Zanini and **Vikas Dhawan** Research Division

Introduction

Recent technological advances have led to the availability of new types of observations and measurements that were previously not available and that have fuelled the 'Big Data' trend (Dhawan & Zanini, 2014). Along with standard *structured* forms of data (containing mainly numbers), modern databases include new forms of *unstructured* data comprising words, images, sounds and videos which require new techniques to be exploited and interpreted.

This article focusses on Text Mining (TM), that is a set of statistical and computer science techniques specifically developed to analyse text data. It aims to give a theoretical introduction to TM and to provide some examples of its applications. Text has always been an informative source of insight into a specific field or individuals. However, with the advent of new technologies, text data are also being predominantly used in new forms of communication. New sources of text data are now available, such as text messaging, social media activity, blogs and web searches. The increasing availability of published text, sophisticated technologies and growing interest in organisations in extracting information from text has led to replacing (or at least supplementing) the human effort with automatic systems.

TM can be used for a variety of scopes, ranging from basic descriptions of text content through word counts to more sophisticated uses such as finding links between authors and evaluating the content of scripts (e.g., automated marking of essays).

TM refers to the process of extracting meaningful numeric indices from text. It owes its origin to a combination of various related fields – Data Mining (DM), Artificial Intelligence, Statistics, Database Management, Library Science and Linguistics (Anawis, 2014). Its basic purpose is to process the unstructured information contained in text data in order to make text accessible to various DM statistical algorithms. This could help make text data as informative as standard structured data and allow us to investigate relationships and patterns which would otherwise be extremely difficult, if not impossible, to discover. With TM, information contained in the text can be categorised and clustered with the aim of producing results such as word frequency distribution, pattern recognition and predictive analytics which might not be easily available using standard data (JISC, 2008).

The possibility of analysing text data is recognised as one of the main elements of the Big Data trend (Lohr, 2012) and a leading source of information for data journalism (Rogers, 2011). In recent years, greater understanding of the potential of TM has led government/public authorities and private organisations to play an active role in developing this technology. The National Centre for Text Mining (NaCTeM) was possibly the first publicly-funded TM centre in the world[1], established by the UK's JISC[2] and operated by the University of Manchester (for an introduction to NaCTeM see Ananiadou, 2005). NaCTeM was established in 2004 to provide TM services in response to the requirements of the UK academic community and to provide leadership in its use in learning, teaching, research and administration. The potential of TM has also been recognised elsewhere in the world. For example, in Italy, Cineca (a consortium made up of 54 Italian universities and the Ministry of Education, University and Research) has been using one of the most powerful computers in the world to design and develop information systems and TM solutions for public administration, health care and business.

TM can be a strategic source of evidence-based information that can support the decision-making process in different fields, from policy-making to business. For this reason, researchers and practitioners from various fields are using TM.

The logic (and technology) behind Text Mining

Broadly speaking, the overarching goal of TM is to turn text into data so that it is suitable for analysis. To achieve this there is a need for applying computationally-intensive artificial intelligence algorithms and statistical techniques to text documents. As stated in a JISC briefing paper (JISC, 2008), TM employs a wide range of tasks that can be combined together into a single workflow, in which it is possible to distinguish four different stages:

1. Information Retrieval
2. Natural Language Processing (NLP)
3. Information Extraction and
4. Data Mining.

Information retrieval

The first stage of TM is to identify the relevant documents from a large collection of digital text documents. Information Retrieval systems used are aimed at identifying the subset of documents which match a user's query. Two examples of Information Retrieval systems are the tools used in libraries to search for books on a specific topic and web search engines (e.g., Google, Bing) designed to search for information in the World Wide Web.

Natural Language Processing

Once a subset of text documents has been retrieved the character strings have to be processed in order to be analysed by computers. The computers need to be fed input in a specific format so that they can understand natural languages as humans do (Manning & Schütze, 1999).

1. See NaCTeM web page at http://www.nactem.ac.uk/
2. JISC (formerly known as the Joint Information Systems Committee) is a UK non-departmental public body whose role is to support post-compulsory education and research, providing leadership in the use of ICT in learning, teaching, research and administration.

The main difficulty is that, often, the hidden structure of natural language is highly ambiguous. Although this might jeopardise the outcome, developments in NLP have led to a high degree of success in certain tasks. NLP enables us to (JISC, 2008):

- classify words into grammatical categories (e.g., nouns, verbs);
- disambiguate the meaning of a word, among the multiple meanings that it could have, on the grounds of the content of the document;
- parse a sentence, that is, perform a grammatical analysis that enables us to generate a complete representation of the grammatical structure of a sentence, not just identify the main grammatical elements in a sentence.

During this stage of TM, the linguistic data about text are extracted from, and marked-up to, the documents which still hold an unstructured form of data.

Information Extraction

In order to be mined as any other kind of data, the unstructured natural language document must be turned into data in a structured form. This stage is called Information Extraction and it is the data generated by NLP systems. The most common task performed during this stage is the identification of specific terms, which may consist of one or more words, as in the case of scientific research documents containing many complex multi-word terms.

Information extraction also allows us to link names and entities (e.g., people and the organisation to which they are affiliated) and more complex facts such as relationships between events or names.

Data Mining

When the structured database is filled with the information extracted from the annotated documents provided by NLP algorithms, data are finally ready to be mined. In this context 'mining' is a synonym of 'analysing', as the aim is to draw useful information from the text data in order to build up new knowledge. To do this, given that data are now in a structured form, it is possible to make use of standard statistical procedures and techniques applied to text data that are now in structured form.[3]

Applications of Text Mining

The first applications of TM surfaced in the mid-1980s.[4] However its growth has been led by technological advances in the last ten years. TM has being increasingly employed in applied research in different areas (such as epidemiology, economics and education) as well as for business-related purposes, especially for gaining market and consumer insights and to develop new products. The techniques of TM are common to both academic research and business-oriented analytics.

3. Among the most common statistical packages used by researchers, the text analytics tools are 'Text Miner' and 'Enterprise Miner' (SAS), 'TM – Text Mining Infrastructure' (R) and 'Modeler' (SPSS).

4. See, for example, the Content Analysis of Verbatim Explanations Research project. http://www.ppc.sas. upenn.edu/cave.htm

From basic word counts to sentiment analyses

Some of the applications of TM require very basic statistics, frequencies for instance. Counting the occurrence of one or more words from a document is the most common TM application, but it does require new ways to visualise this kind of data. For example, Wordle, a free tool available online (http://www.wordle.net/) generates tag clouds of the words contained in a document (Feinberg, 2010). The size of each word is proportional to its relative frequency in the document (similar to a bubble plot).

The technological advances that have fuelled TM development have not just inspired new data visualisations, but also stimulated the collection of new 'textbases', such as *Project Gutenberg* and *Google Books*. For instance, digitising and archiving books allows us to calculate the frequency of a word in a book, or in all the books published in a specific year or to visualise the occurrence of certain words over time. For books available in *Google Books*, Figure 1 gives an example of the occurrence of the words 'information' and 'news' in books published during the last century. Whilst the word 'news' appears to have been steadily used by authors over the last century, the word 'information' experienced a notable increase: from about the same level as 'news' in the early 1900s, to six times more than 'news' in the year 2000.

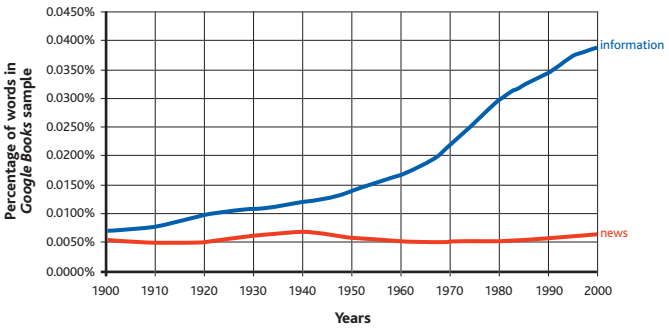

Figure 1: Searches for the words 'information' and 'news' in *Google Books* (digitalised books originally published between 1900 and 2000)

Image sourced from *Google Books Ngram Viewer*.
Retrieved from https://books.google.com/ngrams

Word counts and the availability of large-scale 'textbases' give the opportunity to analyse the evolution of literary styles and trends over time and across countries. This kind of analysis belongs to a new field of study known as 'culturomics' (Ball, 2013). For example, in a recent study, a group of researchers mined a sample of 7,733 works obtained from the *Project Gutenberg* Digital Library written by 537 authors after the year 1550 (Hughes, Foti, Krakauer,& Rockmore , 2012). They focused on the use of 307 content-free words (e.g., prepositions, articles, conjunctions and common nouns) claiming that these words provide a useful stylistic fingerprint for authorship and can be used as a method of comparing author styles. For each author a similarity index with every other author was computed. This index, based on the occurrences of each content-free word considered in the study, was used to exploit temporal trends in the usage of content-free words. Their primary finding was that authors tend to have important stylistic connections to other authors closer to them in time, but not necessarily to immediate contemporaries. They noticed that, for books published within three years of each other, the similarity index is very high, but slightly smaller than the one shown for books published within ten years of each other. For books published with a temporal distance of more than ten years, the similarity index decreased

until reaching a stable value for books published with a temporal distance of 350 years.

Another innovative piece of research, carried out by Matthew Jockers of the University of Nebraska-Lincoln, focused on comparing the stylistic and thematic connections amongst eighteenth and nineteenth century authors. A massive amount of text data using digital versions of nearly 3,500 books was processed to investigate how books were connected to one another on criteria such as frequency of words, choice of words and overarching subject matter (Jockers, 2013). Each book was then affixed with unique attributes and plotted graphically. Figure 2 shows the books analysed from the late 1700s to the early 1900s. The books plotted closer to each other represent a close relationship in terms of styles and themes. Figure 2 highlights the example of Herman Melville's *Moby Dick* published in 1851 which appears here as an outlier from much of the literary work of the period while still being related to several works by James Fenimore Cooper (*Sea Lions* published in 1849 and *The Crater* published in 1847).

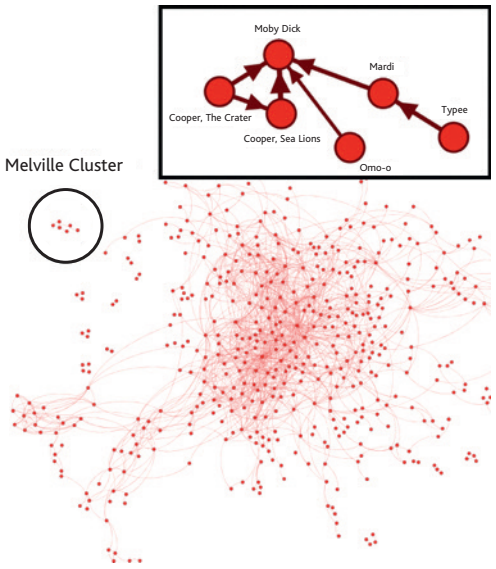

Figure 2: Graphical distribution that displays connections, insights and trends about the literary world from the late 1700s to the late 1900s

Image courtesy of Matthew Jockers (University of Nebraska-Lincoln).

Recent research led by Durham University studied the use of emotion-related words in recent history (Acerbi, Lampos, Garnett & Bentley, 2013). Based on these words this research found that there was a 'sad' peak during the Second World War and two 'happy' peaks – one in the 1920s and another in the 1960s (see Figure 3). A 'sad' period was also noticed during the 1970s and the 1980s followed by an increase in happiness-related words around 1990–2000. The study pointed out that in general, the use of emotion-related words has reduced in the past century. The study also compared historical trends in the use of emotion-related words between British and American authors. Prior to 1980, the difference between them was barely significant, but since then emotion-related words have been used more frequently by US authors than UK ones.

Mining of social opinions is becoming a common marketing and brand management strategy used by organisations. This kind of analysis includes understanding what people say or share in their everyday life, particularly online. This area of research is known as 'opinion mining' or 'sentiment analysis'. Its aim is to identify and extract subjective

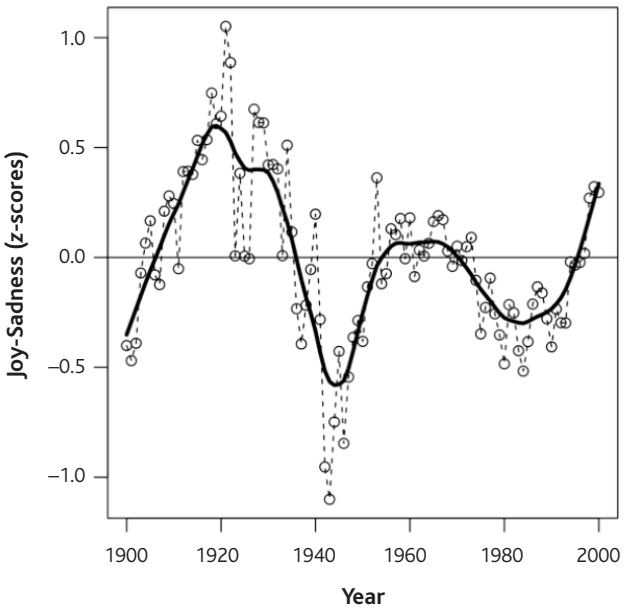

Figure 3: Historical periods of positive and negative moods

Note: Difference between z-scores of Joy and Sadness for years from 1900 to 2000 (raw data and smoothed trend). Values above zero indicate generally 'happy' periods, and values below the zero indicate generally 'sad' periods.

information from text documents such as social media posts. Sentiment analysis is one of the main research strands of Global Pulse, a new initiative by the United Nations (UN) aimed at leveraging the use of Big Data for global development. In a recent work, Twitter conversations related to food price inflation amongst Indonesians were investigated. The research found a significant correlation between official food inflation rates and the number of tweets about this topic (UN, 2014). The study concluded that automated monitoring of public sentiment on social media, combined with contextual knowledge, has the potential to be a valuable real-time alternative to official statistics (usually released after a certain time lag) and to uncover people's reactions in contexts where the use of social media is widespread.

Sentiment mining has also been exploited in other research contexts, such as the understanding of political and historical trends (Ceron, Curini, Iacus & Porro, 2014; Huijnen, Iaan, de Rijke & Pieters, 2014). Social media websites and other computational tools (e.g., *Google Books Ngram Viewer*) are being used for research in this area. This approach could help retrieve hidden information in a large corpus of text documents including speech transcripts by writers and speakers.

Links amongst words and text pattern recognition

Basic statistics are sufficient to summarise, categorise and cluster information from text documents. TM, in addition, may be helpful to generate meaningful links across different documents when decision-makers are overloaded with unstructured information, such as news articles in the case of financial market agents. At times, TM could help reveal unexpected connections between documents. The relationship between the use of certain words in real estate advertisements and the price of the house advertised make an interesting example. In their best-selling book *Freakonomics*, Dubner and Levitt (2005) listed five terms commonly used in real-estate advertisements in the USA associated with

a) higher sale price and b) lower sale price. Table 1 gives the five terms for both (in order of their association with price). The more expensive houses were described using words which were all related to the physical description of the house such as 'granite' and 'maple '. Unexpectedly, words such as 'fantastic' and 'charming' were used more often for cheaper houses. The authors suggest that these words are used as a sort of real-estate agent code to attract potential customers for a house which doesn't have many saleable attributes.

Table 1: Terms used in USA real-estate adverts and their association with house price (Dubner & Levitt, 2005).

Five terms associated with higher price	*Five terms associated with lower price*
Granite	Fantastic
State-of-the art	Spacious
Corian®	!
Maple	Charming
Gourmet	Great neighbourhood

However, we need to be careful in drawing interpretations from text data. For instance, it has been reported in a post written in a language blog by the computational linguistic Mark Liberman[5] that statistically significant correlations were unexpectedly found between words apparently not linked, such as 'some' and 'all', 'the' and 'you'. This suggests that, although it is not hard to find patterns in large datasets, the results may not be meaningful or not always straightforward to interpret and the patterns could also be attributed purely to sampling error.

Word pattern recognition has also been applied to everyday working life. Automated systems (known as eCRM – Customer Relationship Management) have been developed as an attempt to categorise incoming email, and to automatically respond to users with standard answers to frequently asked questions.

One of the most familiar applications of TM technology and machine learning techniques is *Google Translate*, a free, multilingual translation service provided by *Google Inc*. to translate written text from/into 63 languages. *Google Translate* is based on a large scale statistical analysis, rather than traditional grammatical rule-based analysis. To generate a translation, *Google Translate* looks for patterns in hundreds of millions of documents that have already been translated by human translators and are available on the web. This process of seeking patterns in large amounts of text is called 'statistical machine translation' (Och, 2005).[6] Clearly, the more human-translated documents that *Google Translate* can analyse in a specific language, the better the translation quality will be.

Publicly available data and predictive modelling

With the advent of new technologies, a source of data is not just a document for TM, the search for that document itself can provide useful insights. In the case of documents available online, web searches through search engines can be informative. *Google*, for example, set up *Google Trends*, which allows internet users to easily access metrics on *Google* searches.

An example of such trends is given in Figure 4. It shows the comparison of text searches in *Google* for the terms 'OCR', 'Edexcel' and 'AQA' (the names of three awarding bodies based in England, Wales and Northern Ireland) from January 2011 to September 2014.[7] The searches for the three awarding bodies follow a similar pattern to each other which, not unexpectedly, depict a seasonal component: the two peaks are in June and January of year each (except for January 2014[8]), when the majority of students sit the exams, whilst August has fewer searches, when schools are closed. During examination sessions AQA was the most searched, while OCR had the highest number of searches from September to December.[9] *Google Trends* also provides a list of related searches, that is, popular search terms that are associated with the term searched. In the example given here, for all three awarding bodies, the most related search was their name followed by the term 'past papers' (e.g., 'OCR past papers'). The second most frequent related search was the name of the awarding body followed by 'GCSE' (e.g., OCR GCSE). We also observed that while the most searched subject for OCR and AQA was Biology, for Edexcel it was Mathematics.

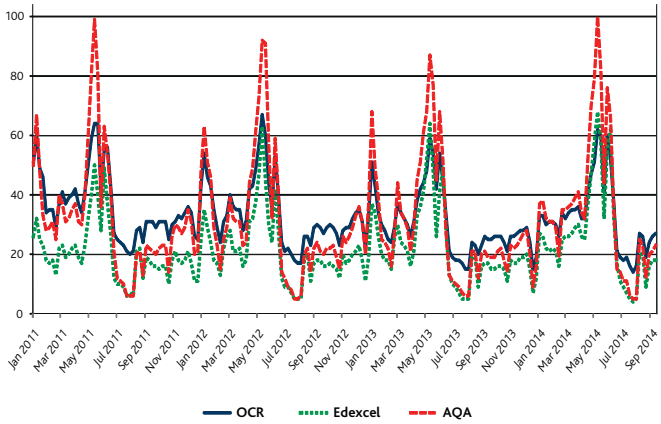

Figure 4: Text searches for 'OCR', 'Edexcel', and 'AQA' from January 2011 to September 2014

Image sourced from *Google Trends*. Retrieved from http://www.google.com/trends

It has been shown that the number of text queries that users enter into web search engines such as *Google* and *Yahoo* can be used for predictive modelling for forecasting values of a number of measures of interest. Researchers in epidemiology discovered that search requests for terms like 'flu symptoms' and 'flu treatments' were a good predictor of the number of patients who, in the period 2004–2008, required access to USA hospital emergency rooms in the next two weeks (Polgreen, Chen, Pennock, Nelson & Weinstein, 2008; Ginsberg et al., 2009). With reference to 2013, it was reported that these web searches were predicting more than double the proportion of doctor visits for influenza-like illness that were actually recorded. This was probably caused by a change in the *Google* search algorithm (Lazer, Kennedy, King, & Vespignani, 2014). Although this discovery can undermine the suitability of web searches as a predictive method, it has been proven to be a good source of

5. *Significant (?) relationships everywhere*. Language Log. Retrieved from: http://languagelog.ldc.upenn.edu/nll/ ?p=4686#more-4686
6. See also the webpage of the Google Research team at http://research.google.com/pubs/MachineTranslation.html
7. *Google Trends* does not provide data on the access to the website (which is something that *Google Analytics* does, though this is not publicly accessible). So the data plotted in Figure 4 are not 'visits' to the three awarding bodies' websites, but only 'searches'. Moreover, data provided does not show the actual volume of searches, but only an indicator estimated in relation to the maximum value of searches across the comparison which is set to 100.
8. It should be noted that in 2014, there was no January exam sitting.
9. Note that the results might have been different if, for instance, 'Pearson' or 'Pearson Edexcel' had been used instead of 'Edexcel'. Pearson has been the parent company of Edexcel since 2003. In 2010, the legal name of the Edexcel awarding body became Pearson Education Limited (Pearson).

information when combined with traditional sources of data. Web search data combined with official statistics have been extensively used to predict the unemployment rate in different countries such as the US (Ettredge, Gerdes, & Karuga, 2005; D'Amuri & Marcucci, 2010), Germany (Askitas & Zimmermann, 2009) and Israel (Suhoy, 2009). It has also been shown that web search data employed as an explanatory variable, along with the previous historical trends of the dimension of interest, can sensibly improve short-term predictions of other social and economic indicators such as inflation (Guzman, 2011). Therefore, predictive modelling could also enable central banks and other national and international agencies to improve the timing and the accuracy of the policy measures they publish to inform policy makers. It can also be applied to economic metrics for business-related purposes and analysing customer insights.

Evidence has shown that web search queries "…can be useful leading indicators for subsequent consumer purchases in situations where consumers start planning purchases significantly in advance of their actual purchase decision" (Choi & Varian, 2012). For instance, search engine data related to housing search enquiries has been shown to be a more accurate predictor of house sales in the next quarter than the forecasts provided by real estate economists (Wu & Brynjolfsson, 2013). Web search queries have also been successfully employed to improve the predictability of motor vehicle demand and holiday destinations (Choi & Varian, 2012). These are applications of the terms attributed to Choi and Varian's – 'contemporaneous forecasting' or 'nowcasting', because they can help in 'predicting the present', rather than the future (Choi & Varian, 2012).

The use of predictive modelling has also been adapted by online retailers to gain customer insights. Amazon and Netflix recommendations, for example, rely on predictive models of what book or film a customer might want to purchase on the basis of their history of enquiries to the website or similar purchases made by other customers (Einav & Levin, 2014). In general, online advertising and marketing tends to rely on automated predictive algorithms that target customers who might be interested in responding to offers.

Predictive modelling based on text data extends well beyond the online world. One of the most famous applications is the development of algorithms that make use of text data contained in different forms of communication (e.g., mobile texts and emails) to detect terrorist threats and to identify fraudulent behaviour in healthcare and financial services (Einav & Levin, 2014).

Applications of Text Mining in education

The benefits offered by the interaction of text and other data analytics in improving learning processes are already being valued by education practitioners as well as by learners themselves.

The first example is the implementation of an experimental real-time case study in a business course. Lecturers made use of internet-based software to facilitate written communication among students, teachers and the case organisation. In this way, it was possible to gather a large quantity of text data containing all the email communication among students and the organisation involved in the case study. Applying simple text analytics on real-time written communication, such as counting of specific words, researchers found that, by the end of this experimental teaching approach, students had increased their understanding of a live business problem. Furthermore, from the analysis of text data, it was possible to discover that, during the case study, students learnt how to use a language more similar to the one used in the real business world. In an evaluation of this experiment, students affirmed that they liked this new teaching approach and would like to see more of it at their schools as they found it very applicable to real life (Theroux, 2009).

A second example of the use of TM to gather insights on learners' cognition is a study aimed at analysing students' progression in a computer programming class. In this study, a software package was used to gather data during a programming assignment from nine learners (Blikstein, 2011). The software allowed researchers to build a 1.5 GB dataset of 18 million lines of events (such as keystrokes, code changes, error messages and actual coding snapshots). An in-depth automated exploration of each student's coding strategies summarised by this mixture of structured and text data was compared with those of other students. The author discovered that error rates progressed in an 'inverse parabolic shape'. This means that, initially, students made a lot of mistakes, but they demonstrated that they were able to learn from them through problem-solving and progressed until they had completed their assignment. Although this is a small-scale study and it is not possible to make any claims about statistical significance, it suggests that using a sophisticated TM application might lead to a better understanding of students' coding styles and sophisticated skills such as problem-solving.

An extensive use of the recent developments in NLP has also been employed to automatically detect secondary students' mental models in order to gain a better understanding of their learning processes. In an experiment students were asked to write short paragraphs about the human circulatory system in order to recall knowledge about the topic. Using an intelligent tutoring system (*MetaTutor*) that teaches students self-regulatory processes during learning of complex Science topics and applying TM techniques, researchers explored which particular machine learning algorithm would enable them to accurately classify each student in terms of their content knowledge (Rus & Azevedo, 2009). Mental models represent an expanding field of research among cognitive psychologists and are aimed at better understanding how well an individual organises content in meaningful ways. TM allows researchers to undertake analysis that can reveal inaccuracies and omissions that are crucial for deep understanding and application of course material, thus informing improvements in course design.[10]

A number of systems using TM have been developed for automated marking of essays and short, free text responses (for an example of the latter see Sukkarieh et al., 2003). Some of the most widely used automated essay marking systems available in the market include: Project Essay Grader, Intelligent Essay Assessor, E-rater, Criterion, IntelliMetric, MY Access and Bayesian Essay Test Scoring System. They have been developed to reduce time and cost and improve reliability and generalisability of the process of assessment in low-stakes classroom tests, as well as for large-scale assessment such as national standardised examinations. The accuracy and reliability of these automated systems have been investigated by educational researchers in the last fifteen years. Along with the benefits of using TM, some of its disadvantages such as the lack of human interaction and the need for a large corpus of sample texts to train the system, have also been reported (Dikli, 2006). Automated essay marking systems do not really understand the texts as

10. For more details on mental model assessment in education see http://mentalmodelassessment.org/

humans do, so it is not possible to affirm that they emulate the human marking process. Notwithstanding, automated essay marking systems show high agreement rates with human markers; and their supporters advocate that the main role of these systems today is not to replace teachers and assessors, but to assist them, incorporating these systems as a supplementary marker, especially in large-scale writing assessments (Monaghan & Bridgeman, 2005; Kersting, Sherin & Stigler, 2014).

A particular example of automated essay marking is the tool developed by a team of researchers at Maastricht University to stimulate students to become active and collaborative learners. It has been used in Statistics courses to assess students on their understanding of course content. It makes use of advanced NLP and Latent Semantic Analysis algorithms that can be used in automatic marking of the texts. Mining students' essays, researchers were easily able to automatically discriminate between the reference book chapter text and the documents of the students. However, it is less clear whether this tool is able to discriminate students from one another (Imbos & Ambergen, 2010).

Despite its weaknesses, marking essays automatically continues to attract the attention of schools, universities, assessment organisations, researchers and educators. Although it might be difficult for these systems to supersede human markers, TM can be employed to support human markers as a second or third marker (see, for instance, Landauer, 2003 and Attali & Burstein, 2006). The Centre for Digital Education (CDE) reported that, in the USA, around $20 billion was spent on public education in Information Technology in 2012, with an increase of 2 per cent from the previous year[11]. The awareness of the potential of TM and DM in, for instance, formative assessment, has led McGraw-Hill to develop two different tools, *Acuity Predictive Assessment* and *Acuity Diagnostic Assessment*, aimed at informing teachers and learners about their performance and how to improve it (CDE, 2014).

These tools can be employed for formative assessment. Predictive modelling of text data can provide an early indication of how students will perform on a standardised test. It allows assessment of the gap between what students are *expected* to know and what they *actually* know. It can also provide evidence regarding which area of the syllabus they have to focus on to improve their performance (West, 2012). Also, more advanced analysis could be informative to teachers about which particular teaching techniques are more efficient for specific students and the best ways to tailor the learning approach to them (Bienkowski, Feng & Means, 2012).

Students' reading comprehension, for example, has been the object of a study based on the use of intelligent tutoring software. The analysis of data such as students' reading mistakes and word knowledge gathered through a speech recognition tool showed that re-reading an old story helped pupils learn half as many words as reading a new story (Beck & Mostow, 2008). An online tool called *WebQuest* provides activities designed for teachers to train pupils in skills such as information acquisition and evaluation of online materials. Students who have experienced these kinds of activities have reportedly enjoyed the collaborative and interactive nature of the activities (Perkins & McKnight, 2005).

Predictive modelling in educational assessment has been mainly based on numeric data (e.g., days of truancy, overall grades and disciplinary problems). However, text data could be used to enable more in-depth analyses in order to get better insights on assessment. For example, Worsley & Blikstein (2011) examined students' dialogues along with other qualitative and quantitative data to develop predictors for student expertise in the area of Engineering design. By leveraging the tools of machine learning, NLP, speech analysis and sentiment extraction, the authors identified a number of distinguishing factors of learners at different levels of expertise. According to the study, these kinds of findings motivate further research in this field and the development of a new paradigm for the evaluation of learner knowledge construction.

11. Centre for Digital Education: http://www.centerdigitaled.com/research/

Discussion

The key advantage provided by TM is the opportunity to exploit text records, on a very large scale. In this article we have briefly described the techniques of TM and some of its applications.

TM has a variety of potential applications in the field of education. In formative and summative assessment, for instance, it could be used to understand trends in vocabulary usage over time and the use of spelling and punctuation. To date, these applications have been carried out by teachers and assessment experts without using advanced techniques such as TM, but TM allows the possibility of implementing these applications on a more comprehensive scale. The developments in NLP allow educational professionals to analyse the language structure of a vast amount of text documents in just a few minutes, plus the ongoing developments in this field could result in an increase in the accuracy of the findings.

The availability of novel data could lead, at least in principle, to novel measurement and research designs to address old and new research questions. However, working with very large, rich and new kind of datasets, it might not be straightforward to figure out what questions the data could answer accurately. Asking the right question might be more important now than ever (Einav & Levin, 2014). Exploiting large text datasets without a proper research question might lead to a significant waste of resources.

More heterogeneous and in-depth data could allow researchers to move from methods that allow the estimation of average relationships in the population towards differential effects for specific subpopulations of interest. This could mean looking at particular categories of students, defined by their specific background, level of achievement and other characteristics of interest. TM is an expanding field with the potential to support innovative areas of research. With careful research designs and proper methods, TM could make a salient contribution to educational research.

References

Acerbi, A., Lampos, V., Garnett, P., & Bentley, R. A. (2013). The expression of emotions in 20th century books. *PloS one*, *8*(3), e59030.

Ananiadou, S., Chruszcz, J., Keane, J., McNaught, J., & Watry, P. (2005). The National Centre for Text Mining: Aims and Objectives. *Ariadne*, *42*. Retrieved from: http://www.ariadne.ac.uk/issue42/ananiadou.

Anawis, M. (2014). Text Mining: The Next Data Frontier. *Scientific Computing*. Retrieved from: http://www.scientificcomputing.com/blogs/2014/01/text-mining-next-data-frontier.

Askitas, N., & Zimmerman, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly* (formerly: *Konjunkturpolitik*), Duncker & Humblot, Berlin, *55*(2), 107–120.

Attali, Y. & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning, and Assessment*, *4*(3).

Ettredge, M., Gerdes, J., & Karuga. G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM, 48*(11), 87–92.

Ball, P. (2013, 21 March). Text mining uncovers British reserve and US emotion. *Nature*. Retrieved from: http://www.nature.com/news/text-mining-uncovers-british-reserve-and-us-emotion-1.12642.

Beck, J., & Mostow, J. (2008). How Who Should Practice: Using Learning Decomposition to Evaluate the Efficacy of Different Types of Practice for Different Types of Students. In B. Woolf, E. Aïmeur, R. Nkambou & S. Lajoie (Eds.), *Intelligent Tutoring Systems,* (5091), 353–362. Springer Berlin Heidelberg.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief*. U.S. Department of Educational, Office of Educational Technology. Retrieved from: http://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf

Blikstein, P. (2011). *Using learning analytics to assess students' behavior in open-ended programming tasks*. Paper presented at the Proceedings of the 1st international conference on learning analytics and knowledge.

Centre for Digital Education (CDE) (2013). Big Data, Big Expectations. *The Promise and Practicability of Big Data for Education*. The Centre for Digital Education. Retrieved from: http://www.centerdigitaled.com/paper/259374351.html

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New media and society*, *16*(2), 340–358.

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, *88*(1), 2–9.

D'Amuri, F., & Marcucci, J. (2010). *"Google it!" Forecasting the US unemployment rate with a Google job search index*. ISER Working Paper Series 2009–32. Institute for Social & Economic Research (ISER).

Dhawan, V., & Zanini, N. (2014). Big data and social media analytics. *Research Matters: A Cambridge Assessment Publication*, *18*, 36–41.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1).

Dubner, S. J., & Levitt, S. D. (2005). *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. New York City: William Morrow.

Einav, L., & Levin, J. D. (2014). The Data Revolution and Economic Analysis. *Innovation Policy and the Economy*, *14*(1), 1–24.

Feinberg, J. (2010). Wordle, in J. Steele & N. Iliinsky (Eds.) *Beautiful visualization*, Sebastopol: O'Reilly Media, Inc.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014.

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement*, *36*(3), 119–167.

Hughes, J.M., Foti, N. J., Krakauer, D. C., & Rockmore D. N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences*, *109*(20), 7682–7686.

Huijnen, P., Laan, F., de Rijke, M., & Pieters, T. (2014). A Digital Humanities Approach to the History of Science. In A. Nadamoto, A. Jatowt, A. Wierzbicki & J. Leidner (Eds.), *Social Informatics,* (8359), 71–85. Springer Berlin Heidelberg.

Imbos, T., & Ambergen, T. (2010). *Text analytic tools for the cognitive diagnosis of student writings*. Paper presented at the Proceedings of the ICOTS8, International Conference on Teaching Statistics.

JISC (2008). *Text Mining Briefing Paper*. Joint Information Systems Committee. Retrieved from: http://jisc.ac.uk/media/documents/publications/bptextminingv2.pdf.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.

Kersting, N. B., Sherin, B. L. & Stigler, J. W. (2014). Automated Scoring of Teachers' Open-Ended Responses to Video Prompts: Bringing the Classroom-Video-Analysis Assessment to Scale. *Educational and Psychological Measurement*, *74*(6), 950–974.

Landauer, T. K. (2003). Automatic Essay Assessment, Assessment. *Education: Principles, Policy & Practice*, *10*(3), 295–308.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014, 14 March). The Parable of Google Flu: Traps in Big Data Analyisis. *Science*, *343*(6176), 1203–1205. Retrieved from: http://www.sciencemag.org/content/343/6176/1203.

Lohr, S. (2012, 11 February). The Age of Big Data. *The New York Times*. Retrieved from: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0.

Manning, C.D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Boston: MIT press.

Monaghan, W., & Bridgeman, B. (2005). E-rater as a Quality Control on Human Scorer. *ETS RD Connections*. Retrieved from: http://www.ets.org/Media/Research/pdf/RD_Connections2.pdf.

Och, F.J. (2005). *Statistical Machine Translation: Foundations and Recent Advances*. Retrieved from: http://www.mt-archive.info/MTS-2005-Och.pdf.

Perkins, R., & McKnight, M.L. (2005). Teachers' attitudes toward WebQuests as a method of teaching. *Computers in the Schools*, *22*(1–2), 123–133.

Polgreen, P.M., Chen, Y., Pennock, D. M., Nelson, F. D., & Weinstein, R. A. (2008). Using Internet Searches for Influenza Surveillance. *Clinical infectious diseases*, *47*(11), 1443–1448.

Rogers, S. (2011, 28 July). Data journalism at the Guardian: what is it and how do we do it? *The Guardian Datablog*. Retrieved from: http://www.theguardian.com/news/datablog/2011/jul/28/data-journalism

Rus, V., Lintean, M., & Azevedo, R. (2009). *Automatic Detection of Student Mental Models during Prior Knowledge Activation in MetaTutor*. International Working Group on Educational Data Mining. Paper presented at the International Conference on Educational Data Mining (EDM) (2nd, Cordoba, Spain, July 1–3, 2009).

Sukkarieh, J. Z., Pulman, S. G. & Raikes, N. (2003). *Auto-marking: using computational linguistics to score short, free-text responses*. Paper presented at the Proceedings of 29th International Association for Educational Assessment (IAEA) Annual Conference.

Suhoy, T. (2009). *Query indices and a 2008 downturn: Israeli data*. Discussion paper No. 2009.06. Research Department, Bank of Israel.

Theroux, J.M. (2009). Real-time case method: analysis of a second implementation. *Journal of Education for Business*, *84*(6), 367–373.

United Nations (UN) (2014). *Mining Indonesian Tweets to Understand Food Price Crises*. UN Global Pulse Report. Retrieved from: http://www.unglobalpulse.org/sites/default/files/Global-Pulse-Mining-Indonesian-Tweets-Food-Price-Crises%20copy.pdf.

West, D.M. (2012). *Big Data for Education: Data Mining, Data Analytics, and Web Dashboards*. Retrieved from: http://www.brookings.edu/research/papers/2012/09/04-education-technology-west.

Worsley, M., & Blikstein, P. (2011). *Using machine learning to examine learner's engineering expertise using speech, text, and sketch analysis*, in Paper presented at the 41st Annual Meeting of the Jean Piaget Society (JPS). University of California, Berkeley.

Wu, L., & Brynjolfsson, E. (2013). The future of prediction: How Google searches foreshadow housing prices and sales, in S. M. Greenstein, A. Goldfarb and C. Tucker (Eds.) *Economics of Digitization*, Chicago: University of Chicago Press.