Topic modelling for open-ended survey responses

Received (in revised form): 15th February, 2018

Song Chen

is an assistant professor at University of Wisconsin — La Crosse, with a PhD in applied mathematics. He conducts research in scientific computing and data mining with applications in various fields, including marketing. He is director of the data science group at University of Wisconsin — La Crosse and actively leads undergraduate collaborative projects with local industry.

1007 Cowley Hall, Department of Department of Mathematics and Statistics, University of Wisconsin — La Crosse, La Crosse, WI 54601, USA
Tel: +1 608 785 8826; E-mail: schen@uwlax.edu

Chad Vidden

is an assistant professor at the University of Wisconsin — La Crosse, where he leads a data science and mathematical modelling research group that collaborates with local companies. He has a PhD in applied mathematics, with expertise in computational mathematics, data science and machine learning.

1009 Cowley Hall, Department of Mathematics and Statistics, University of Wisconsin — La Crosse, La Crosse, WI 54601, USA
Tel: +1 608 785 5214; E-mail: cvidden@uwlax.edu

Nicole Nelson

is an analytical manager at Kwantum LLC, where she assists the Chief Data Scientist to conduct analytical projects and deliver the business solutions to clients. Nicole's experience is specialised in key-driver modelling, marketing segmentation, Maxdiff analysis, data visualisation and text analytics. She has a chemistry degree and mathematics minor from University of Wisconsin — La Crosse.

Kwantum LLC., 119 19th St. N. La Crosse, WI 54601, USA
E-mail: nicole.hanako@gmail.com

Marco Vriens

is CEO of Kwantum LLC and faculty member at the University of Wisconsin — La Crosse. He has a PhD in marketing analytics, and is an expert in applied analytics. He has led analytics teams for Microsoft, GE and supplier firms. Marco is the author of three books: 'The Insights Advantage: Knowing How to Win' (2012), 'Handbook of Marketing Research' (2006) and 'Conjoint Analysis in Marketing' (1995). His work has been published in academic and industry journals and he has won several best paper awards including the David K. Hardin Award.

Department of Marketing, University of Wisconsin — La Crosse, La Crosse, WI 54601, USA
E-mail: mvriens@uwlax.edu

Abstract Due to the availability of massive amounts of text data, both from online (Twitter, Facebook, online forums, etc) and offline open-ended survey questions, text analytics is growing in marketing research and analytics. Most companies are now using open-ended survey questions to solicit customer opinions on any number of topics (eg 'how can we improve our service?'). With large sample sizes, however, the task of collating this information manually is practically impossible. This paper describes an end-to-end process to extract insight from text survey data via topic modelling. A case study from a Fortune 500 firm is used to illustrate the process.

KEYWORDS: text analysis, open-ended questions, topic modelling, latent Dirichlet allocation, natural language processing

INTRODUCTION

Text analytics applies algorithms to process text data. Due to the massive amount of such data, text analytics is becoming a fast-growing field.[1–3] However, the number of practical papers in marketing research and marketing analytics is still relatively sparse. Many companies have access to text data, such as online data (Facebook, Twitter, forums, blogs, etc) or responses to open-ended survey questions. Businesses still struggle with text analytics because: (1) text analytics models use heavy math theory and computer programming, which can be a challenge for the marketing analyst; (2) there is no uniform model as the task is different for different industries, hence every industry has a different word library, semantics and sentiments (for instance, the text model used for analysing Twitter is totally different from the text model for Yelp reviews); and (3) most problems are not rigidly structured, which makes formalising the process even harder. As a result, there is currently no low-cost, well-defined process that is easily adaptable for companies that are new to this area.

Responses to open-ended survey questions represent one source of unstructured text data with valuable business information. There are, of course, different types of open-ended questions. Popular open-ended survey questions ask why a respondent feels a certain way (eg the main reason for their dissatisfaction) or if they have suggestions to improve the product or service. Customers can voice their opinions freely on any topic they choose. The idea of extracting business insight from such questions is very appealing because such text data truly represent the voice of the customer and may yield unexpected insights. However, for large sample sizes, extracting business insight is practically impossible to do manually, and any firm who wants to do this will face the following trade-offs:

- *Manual vs automation:* The more it leans on the manual side, the better it captures the semantic meanings of the text. On the other hand, manual interpretation is costly and slow, and hence automated ways of analysing text data need to be used at the cost of losing the precision of human interpretation.
- *All-around vs specific:* Software can provide an all-around solution with less sophisticated models while deep learning[4] can achieve more specific goals.

This paper defines a process which combines an automated algorithm and minimal human interpretation.

The text analytics process described is designed for open-ended survey response analysis but is sufficiently flexible to be used for other types of unstructured data with minor modifications. The idea will be illustrated through a (disguised) case study recently conducted for a Fortune 500 company that wanted to mine useful information from 11,000 responses to the open-ended survey question: 'why do you give our service the rating you gave'? Three business questions are addressed:

- What topics are people talking about and is it possible to identify and describe them efficiently? (Once topics have been identified, it is straightforward to look at differences among different business groups.
- How can the results be made actionable?
- Is it possible to build a typing tool to classify new comments into topics for future data?

The paper is organised as follows. First, it describes a detailed process that tackles these three business problems. Secondly, it investigates how to reduce the manual human component. The discussion offers some suggestions for future improvements and extensions.

TEXT ANALYTICS PROCESS

Figure 1 outlines the needed workflow in six major steps: (1) cleaning and preparation;

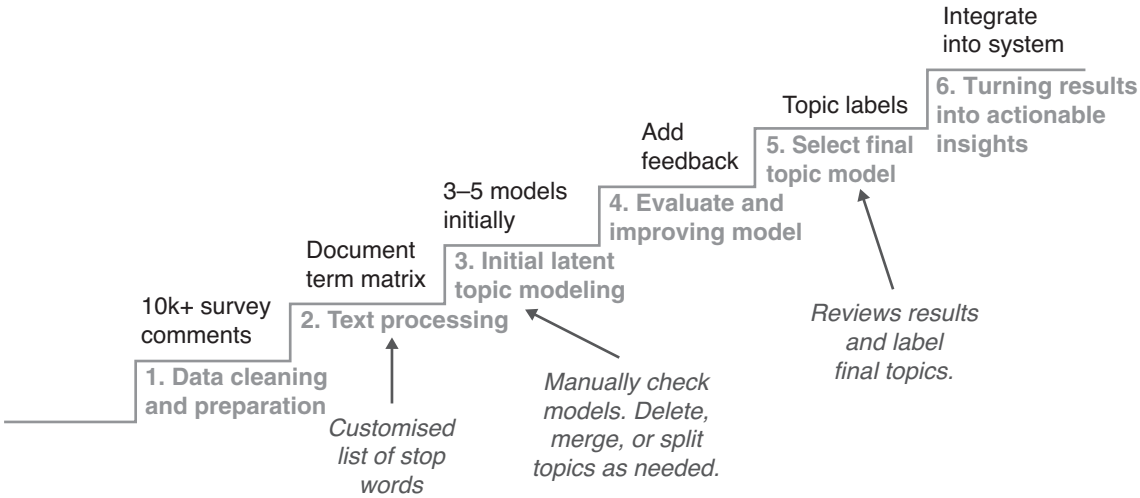

Figure 1: The text analytics process

(2) creating the document to term matrix; (3) running initial latent topic models; (4) evaluating, pruning and rerunning a smaller set of topics models; (5) selecting a final topic model; and (6) making the results actionable and develop a typing tool. The cursive remarks are the parts that require inputs from human experts.

Cleaning and preparation

First, the data must be collected. In the present case, more than 11,000 users replied to the question posed. It is helpful here that the question was phrased in a rather direct way (ie 'How can we improve?'). This allowed for all respondents to approach the question in a similar way and thus using words and sentiment in a consistent manner. The firm collects such data continuously.

Next, the text must be processed to isolate the most valuable information and identify repeated meaning across replies. This is a standard procedure that can be found in any text analytics book[5] including sentence detection, spell check, tokenisation, stemming and lemmatisation, stop words deletion, disambiguation, part of speech tagging, entity extraction and n_gram extraction. Table 1 provides an overview of steps. The goal is to break each sentence into a bag of keywords.

Note that the dimension of keywords quickly becomes very large, so it is recommended to put some restrictions on the keywords selection, for example:

- to avoid singular comments, select only those keywords that appear in more than 2 per cent of the comments;
- to avoid general comments like 'great', 'nothing', keywords need to be mentioned in less than 50 per cent of the comments; and
- deletion of stop words should be done in a way custom to the application at hand. That is, one should delete standard stop words such as 'the' and 'is' but also ones which carry little meaning for the question analysed. For the present study data, phrases like 'would like to see' and 'feel like' were considered stop words and deleted. Without this, topics surrounding such phrases dominate the identified topics later in the process.

In the end, the text document (ie the open-ended responses from the 11,000 users) was reduced to 2,634 unique words. The process can also be improved by using a

Table 1: Descriptions of standard text processing steps

Text processing step	Description
Spell check	Correct any spelling mistakes.
Tokenisation	Isolate individual words from text.
Lemmatisation	Group words with equivalent meaning and rename so all are the same. Example: 'run' and 'ran' would both be renamed 'run'.
Stemming	Reduced versions of the same word to its word stem. Example: 'run', 'running' and 'runs' are all renamed 'run'.
Deleting stop words	Delete words which do not add information to the text. Examples: 'the', 'a', 'am'
Disambiguation and part of speech tagging	Identify words that have multiple meanings according to the context. Example: Mr Green wears a green shirt. One 'green' means name and the other means colour.
Entity extraction	Recognise words such as name, location, or organisations.
N-gram detection	Combine words which are commonly grouped together into a single term. Examples: 'profile_views' and 'job_opportunities'

customised dictionary designed specifically for the business domain. For instance, words such as 'factory warranty' are very important to a manufacturer but less relevant to a retailer. Similarly, each company may have a dictionary of words that works specifically for its industry, which can be easily adapted in the text cleaning process.

Document to term matrix

After each comment is broken down into several keywords, one can transfer the text comments into a numerical matrix known as the document to term matrix (DTM), as illustrated in Table 2.

Each row is a separate comment and each entry represents the keywords frequency, that is, how many times the keywords were mentioned in that comment (for instance, the first comment mentions two keywords 'Easy' and 'Build profile'). The DTM summarises the information document — keyword relationship in all the comments. The DTM is usually large; in the present case, it has around 11,000 rows (comments) and 2,634 columns (keywords). Comments to open-ended questions usually have a very sparse DTM (with more than 95 per cent entries being zero) because each comment hits only a small pool of all the keywords. Using the DTM structure, the text data were transferred into numerical data ready to be used in the next analysis.

Initial latent topic modelling

Next, comments were grouped into topics via the topic modelling procedure.[6] Topic modelling is an unsupervised learning

Table 2: The document to term matrix

Comments	Easy	Create content	Excellent	Service	Old_Format
Make it easier	1	1	0	0	0
Excellent service	0	0	1	1	0
The old format and service was easier to use	1	0	0	1	1

technique, similar to clustering, and aims to define multiple latent (hidden) topics to summarise massive amounts of customer responses. Several algorithms have been developed and tested.[3,6–8] All state-of-the-art research suggests that the latent Dirichlet allocation (LDA) algorithm[6] is a sophisticated yet efficient statistical model. LDA is based on the frequency of words within responses. Within the context of the present paper, any reference to topic models refers to latent Dirichlet allocation models.

Topic modelling takes the DTM as inputs and generates two outputs. It first clusters all the keywords into multiple topics (in essence, segments consisting of groups of keywords) where each topic consists of keywords that are frequently mentioned together, and then generates a matrix representing the topic–keywords relationship. More specifically, a number of topics are obtained, and under each topic are a number of keywords. Each keyword also has a probability of belonging to that topic. It then connects the document–keywords and keywords–topic relationship to produce a document–topic relationship matrix. Table 3 shows an example of such matrices.

In Table 3, each entry is a probability that the two objects are associated together. For instance, Table 3a shows that topic 1 is closely associated with keywords 'easy' and 'create content' and partially associated with 'service'. In Table 3b, comment 1 refers solely to topic 1 while comment 3 is partially related to both topic 1 and 2 but leaning towards topic 2. Using the information in Table 3, one could define what each topic means by looking at the keywords associated closely to that topic. For instance, topic 1 is probably about building content while topic 2 is about easier access.

If one wants each comment to belong to a unique topic, one can take the topic with the biggest probability (thus comment 3 belongs to topic 2) but there are benefits in allowing comments to belong to multiple topics in general. However, when one generates the model to the original 11,000 comments, one faces the following challenges:

- *How to determine the number of topics?* This is not an easy task as one can easily generate over 70 solutions. Anything from a 20-topic solution to a 70-topic solution can be a potentially viable good solution.
- *Interpreting the topics:* When data get bigger and messy, both the topic-keywords and document-topic matrices become much less clear, with a lot of low probabilities assigned (like 0.01), which indicates a very weak association, and the keywords of each topic get very noisy, making it impossible to define topic meanings just by looking at keywords. It is common to find that the keywords are mixed and have no clear meaning, even if one looks at as many as the top 50 keywords of that topic.

Table 3: Sample matrices (a) topic–keywords matrix (top); (b) document–topic matrix (bottom)

Topic	Easy	Create content	Excellent	Service	Old_Format
Topic 1	0.8	1.0	0.0	0.1	0.0
Topic 2	0.2	0.0	1.0	1.0	1.0

Comments	Topic 1	Topic 2
Make it easier to create content	1.0	0.0
Excellent service	0.0	1.0
The old format and service was easier to use	0.2	0.8

Evaluating and further improving the topic model

Running topic modelling is very similar to running cluster analysis. One starts by running a lot of different solutions (in the present case: from 20 to 70 topic solutions) and then evaluating these solutions using statistical measures and judgment as to how useful the solutions are from the perspective of business actionability.

To start with, statistical metrics are used.[9] This study started with in and out-of-sample fit. However, none of these fit metrics yielded a clear and compelling indication of what solution was best (ie they all achieved more or less the same fit). Next, several approaches were used that have been proposed in the literature.[10–13] It is beyond the scope of this paper to discuss these in detail. Figure 2 shows the results of these four metrics for a range of solutions.

These metrics helped to narrow the choice down to three solutions. Looking at Figure 2, the aim is to pick those topic numbers that minimise the two curves at the top while maximising the two curves at the bottom. The reasonable choices are 42, 51 and 59. These statistical measures do not agree with each other perfectly, so the next step is to pick diverse numbers of topics (instead of picking 51 and 52, which are too close together, 42 and 51, for example, represent a better choice). Human interpreters can achieve the precision in due course.

Turning topic models into actionable insights

Figure 3 shows the selection of a final topic model and the process of deriving actionable insights.

As a first step, each topic solution is manually reviewed by looking at the keywords and top related comments. For example, consider that the result has been narrowed down to a 42-topic, a 51-topic and a 59-topic solution. To help interpret the solution, a threshold (eg 0.3) is set, and comments with all probabilities falling below this threshold, indicating that the comment does not belong to any topic, are deleted. This reduces the data to

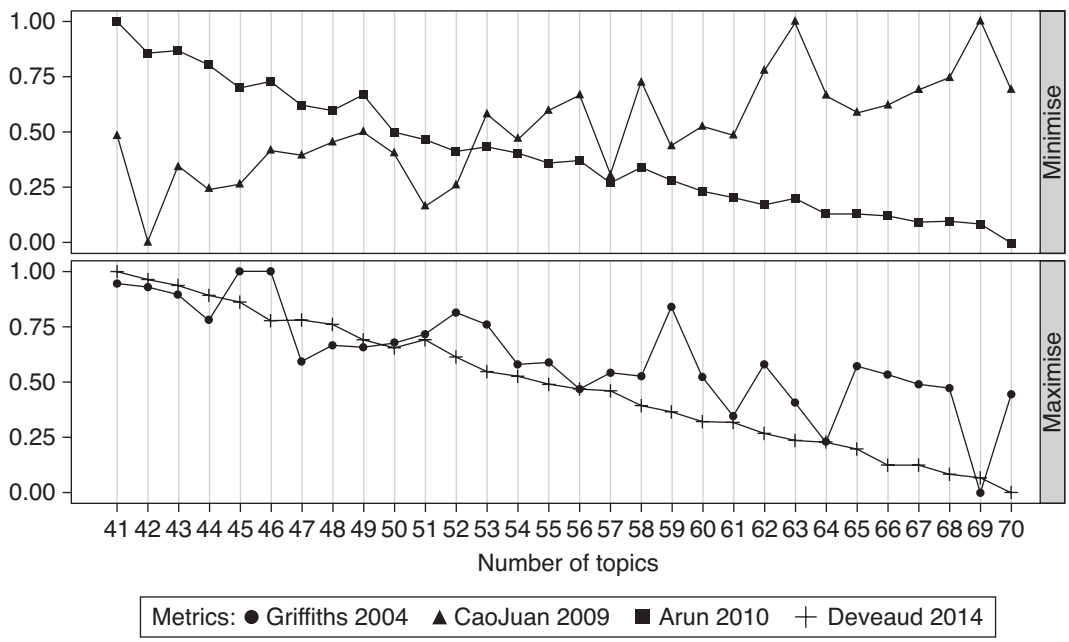

Figure 2: Four standard metrics for assessing LDA topic quality — the top two graphs show optimal topics where values are minimised, while the bottom two show optimal topics where values are maximised

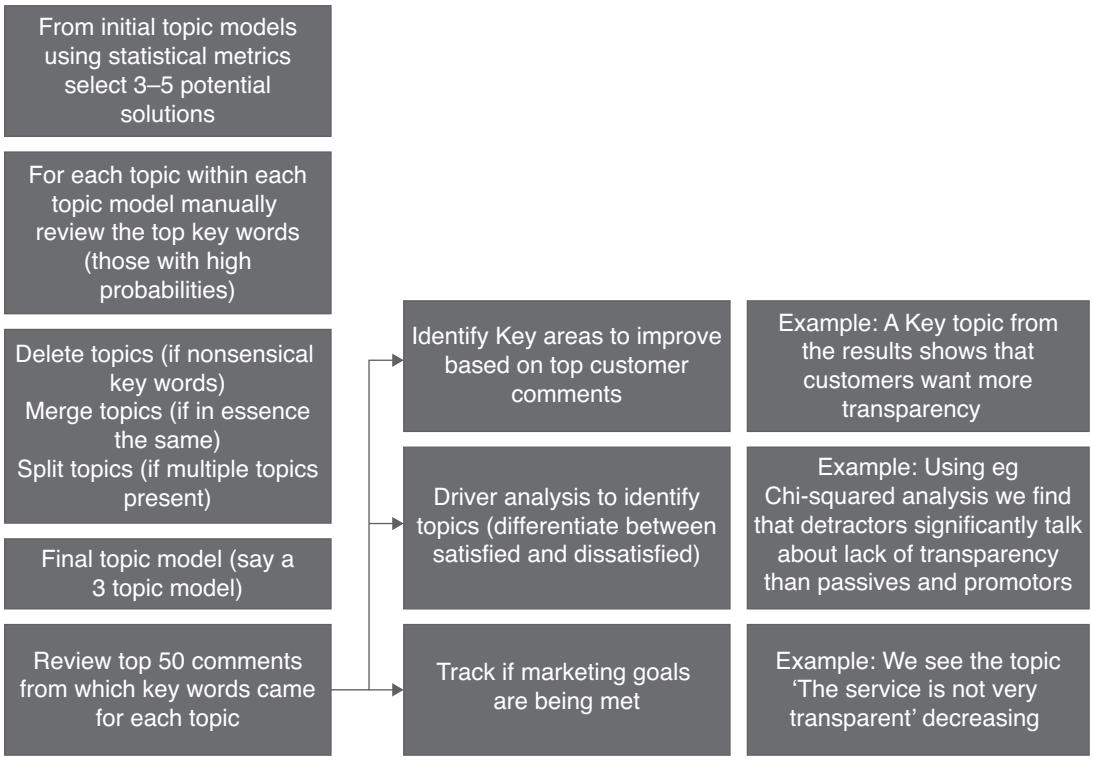

Figure 3: Turning topic models into actionable insights (based on 'What is the one thing you would like to improve?')

20–30 per cent of the original size. This is a big drop, but nevertheless a reasonable approach. In the present instance, most of the comments that were dropped contained little information, eg 'Nothing particular'. Some meaningful comments may well be lost, but if they cannot be classified to any topic then such comments should be treated as the voice of a small minority. Although some information may be lost, most of the high-quality comments are retained with a clearer structure.

For each topic within each solution, the top keywords are reviewed first. If there are any non-meaningful topics, these are marked for deletion. If two topics are in essence the same, they are merged; likewise, if it looks like a topic actually contains multiple topics, it is split.

After modifying the topic solutions (ie each topic is well separated and there are no meaningless topics), the topic solution with the highest quality is selected as the final topic solution. For example, this might be the 51-topic solution (which because of the splitting, merging, deleting, could now be a 50-topic solution).

So, now it is time to perform a more in-depth interpretation. Defining topics based solely on keywords is difficult because one loses the semantic structure in the text. Meanwhile, reading all the comments on each topic is extremely time-consuming. So, after reviewing the top keywords, one turns to the top comments.

Table 3 provides a very simplified illustration as the present study obtained 51 topics; the 'real' Table 3 would be much bigger, and each topic could have multiple comments. The meaning of each topic is defined by reading only the top related comments, that is, the ones with the highest probabilities in the document–topic matrix. The top comments vary from one topic to another. In this way, one retains the semantic structure of the core comments

of each topic while greatly reducing the workload of reading 11,000 comments to reading 20–40 sentences for each topic. To ensure the results match closely, several topics are manually sampled to compare the quality of using all comments and only the top topic related comments (see also next section).

Now one moves to the final step: deriving actionable insights. There are three elements to this part (see Figure 3, right side). First, when reviewing the top comments, one may find suggestions that point to fairly specific actions. For example, a top comment may be that customers want more transparency. Reviewing the top comments will generate a list of potential marketing actions. Next, one can perform 'driver' analysis. That is, topic modelling results can be used for further analysis. Consider the example of a firm that collects likelihood to recommend data and uses this information to calculate a net promoter score (NPS — a popular tool to track how the business is performing).[14] The firm wants to compare what topics come up with respondents who give low likelihood to recommend ratings (referred to as detractors) versus respondents who give higher ratings (passives and promoters). It is possible to test these differences using a chi-square test. One may find, for example, that detractors talk significantly more than other respondents about lack of transparency. By using the chi-squared analysis to examine the top comments, one can narrow the list of actions and translate findings into specific, targeted marketing actions.

The third element of actionability is to develop a tracking system. The probabilities of the topic–word matrix of Table 3 can be used to classify new responses to the same topics on future surveys. Building such a system is called a typing tool. Likelihood to recommend and NPS are often tracked continuously. Now, as new comments come in, they should go through the exact text processing outlined above before topic classification is performed. The typing tool can then be integrated as an automated system. For the present case study, 1,100 new comments from a future quarter were assigned topics via the typing tool. By inspecting the topic assignments for each comment by hand, it was found that 83 per cent were correctly classified. This has an important benefit. Consider, for example, that the firm decided to take measures to improve overall transparency. If they executed this well, then over time the 'transparency' topic should diminish and eventually maybe go away completely. This allows marketers to assess whether their marketing is working.

One drawback of this typing tool model is that new topics cannot be detected. As time goes on and topics change, the topic model may classify fewer comments. It is recommended to create a threshold such that when the certainty of topic classification drops too low, a recommendation for rebuilding the topic model is given.

MAKING HUMAN INTERPRETATION EASIER

There are a few key steps in the above workflow which deserve further discussion. First, steps that involve human interaction play a crucial role in the quality of the final model. As mentioned above, customising the list of stop words is essential, and it is effective to pair this step with an initial exploration of topics. Many stop-word phrases are not apparent until they appear in non-cohesive topics. Secondly, manually inspecting top comments, in a given topic solution, say the 65-topic solution, is time-consuming yet unavoidable when assessing the solution. Inspecting the top 20 comments for 50 or more topics on five separate topic models requires manually reading 5,000 comments.

Defining topics based solely on keywords is difficult because one loses the semantic structure in the text. This paper recommends

Table 4: Comparing three approaches to text summarisation

Summarisation method	Assignment quality (%)	Relative time needed (%)
Five-sentence summary	62	40
50 word summary	83	10
200 word summary	93	20

defining the meaning of each topic by reading only the top related comments. In this way, one retains the semantic structure of the core comments of that topic while greatly reducing the workload of reading 60,000+ comments to reading 20–40 sentences for each topic. To compare the quality of using all the comments and the topic related comments only, a number of topics can be manually sampled. To further reduce the workload, one can apply three text summarisation algorithms to summarise all the comments of a single topic into (1) five sentences, (2) 200 words and (3) 50 words. The performance of each method is shown in Table 4. For each method, the topics are defined by reading the paragraph summarised by the algorithm. One can then assess the result of each method by measuring how much it agrees with the one generated by reading the original comments and how much time it needs (assuming the time takes to read the original comments is 100 per cent).

One can see that summarising to five sentences fails while summarising to 200 words has a 93 per cent hit rate. In practice, it is recommended to use the 200 words approach, but it is possible to use fewer words if one is prepared to sacrifice some accuracy.

DISCUSSION

This work outlines a complete end-to-end process to extract insight from text survey data via topic modelling. The importance of human interaction is emphasised throughout the workflow, integrating domain knowledge into the process and thus maximising the utility of results. The amount of manual interpretation is quite substantial in the topic modelling analyses described. The study has illustrated how actionable results can be derived from the topic model: directly from the comments, further prioritised via chi-square analysis, and infused into a tracking system via a typing tool. Lastly, recommendations are given to improve (reduce) human involvement. The paper tested three ways to make this process less time-consuming, and found that by using 200-word summaries, one can achieve an 80 per cent reduction in the manual time required while still maintaining good accuracy.

This paper provides market researchers with a first step into the realm of text analytics and topic modelling. For academics, this methodology is of value, as it allows them to turn large amounts of unstructured data in to quantitative chunks for further analysis, such as in the development of a new brand measure.[15] For practical, commercial market researchers, this allows them to integrate open-ended responses into the quantitative analyses that are typically done on closed-ended survey questions.

There is more to explore. There are two promising approaches to significantly improve the model in the near future. The topic model used is the latent Dirichlet allocation, which uses the DTM and hence does not capture the semantic meanings of words. Other topic models, such as HMM (Hidden Markov Model)[16] LDA and skip gram methods,[17] should be tested to better mine semantic information from the text. One can then reduce the workload of human experts by increasing the accuracy of the algorithm. Secondly, one can apply dynamic topic modelling,[18] which can track the change of topics over time so that the model will be updated as new comments are received.

References

1. Moe, W., Netzer, O. and Schweidel, D. (2016) 'Social media analytics. Handbook of Marketing Decision Models', in: Wierenga, B. and van der Lans, R. (eds), Springer Science and Business Media.
2. Moreno, A. and Redondo, T. (2016) 'Text analytics: The convergence of Big Data and artificial intelligence', *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 3, No. 6, pp. 57–64.
3. Keating, B.P. (2016) 'Text into numbers: Can marketers benefit from unstructured data', *Applied Marketing Analytics*, Vol. 2, No. 2, pp. 111–120.
4. Schmidhuber, J. (2015) 'Deep learning in neural networks: an overview', *Neural Networks*. Vol. 61, pp. 85–117.
5. Keating, B.P. and Wilson, J.P. (2018) 'Forecasting and Predictive Analytics, Holton Wilson, McGraw-Hill, New York, NY.
6. Blei, D.M., Ng, A. and Jordan, M. (2003) 'Latent Dirichlet allocation', *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.
7. Blei, D.M. (2012) 'Probabilistic topic models', *Communications Association for Computer Machinery (ACM)*, Vol. 55, No. 4, pp. 77–84.
8. Zhang, H., Kim, G. and Xing, E.P. (2015) 'Dynamic topic modeling for monitoring market competition from online text and image data', in Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, pp. 1425–1434.
9. Vidden, C., Vriens, M. and Chen, S. (2016) 'Comparing clustering methods for market segmentation: a simulation study', *Applied Marketing Analytics*, Vol. 2, No. 3, pp. 225–238.
10. Griffiths, T.L. and Steyvers, M. (2004) 'Finding scientific topics', *Proceedings of the National Academy of Sciences*, Vol. 101 (Suppl. 1), pp. 5228–5235.
11. Arun, R., Suresh, V. and Veni Madhavan, C.E. and Narasimha Murthy, M.N. (2010) 'On finding the natural number of topics with latent dirichlet allocation: some observations', in Zaki, M.J., Yu, J.X., Ravindran, B. and Pudi, V. (eds) 'Advances in Knowledge Discovery and Data Mining', Springer, Berlin Heidelberg, pp. 391–402.
12. Juan, C., Tian, X., Jintao, L., Yongdong, Z. and Sheng, T. (2009) 'A density-based method for adaptive LDA model selection', in 'Neurocomputing — 16th European Symposium on Artificial Neural Networks, Bruges', Vol. 72, No. 7–9, pp. 1775–1781.
13. Deveaud, R., SanJuan, E. and Bellot, P. (2014) 'Accurate and effective latent concept modeling for ad hoc information retrieval', *Document numérique*, Vol. 17, No. 1, pp. 61–84.
14. Reichfeld, F.F. (2003) 'The one number you need to grow', *Harvard Business Review*, Vol. 88, No. 12, pp. 46–54
15. Vriens, M., Chen, S. and Ho, C. (2018) 'The evaluation of a new brand measure based on open-ended responses', working paper.
16. Wallach, H. (2006) 'Topic modeling: beyond bag of words', in Proceedings of the 23rd International Conference on Machine Learning, Pittsburg, PA, 25–29th June.
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) 'Distributed representations of words and phrases and their compositionality', in Burges, C.J. C., Bottou, L., Welling, M., Ghahramani, Z. and Weinberger K.Q. (eds) 'Proceedings of the 26th International Conference on Neural Information Processing Systems — Vol. 2', Curran Associates Inc., Lake Tahoe, NE, pp. 3111–3119.
18. Blei, D. and Lafferty, J. (2006) 'Dynamic topic models', in International Conference on Machine Learning, ACM, New York, NY, pp. 113–120.