Schiff Consulting

Introduction to automated text analysis for research and evaluation

Version 1.0, 14 October 2022

Aaron Schiff (aaron@schiff.nz)

Summary

This paper gives an overview of three methods for automated analysis of text data that can be used in research and evaluation work: **sentiment analysis**, **topic detection**, and **summarisation**. The focus is on what is possible rather than technical details, and each method is illustrated in the context of a real-world research or evaluation task.

Automated text analysis cannot entirely replace human analysis and synthesis but can be used alongside manual review to improve productivity and accuracy of analysing text data. Limitations and caveats for using automated text analysis are discussed so evaluators and researchers can gain a realistic understanding of how these methods can be used appropriately and effectively in their work.

Overview of automated text analysis methods for research and evaluation

	What it does	Example applications in research and evaluation
Sentiment analysis	Classify sentiments expressed in short blocks of text as positive, neutral, or negative	• Classifying and ranking responses to open-ended survey questions • Summarising interview responses to specific questions • Processing customer or client feedback or reviews • Extracting data from social media posts
Topic detection	Assess whether blocks of text relate to pre-specified topics *or* Automatically detect the common topics referred to in multiple blocks of text	• Classifying data subjects into groups based on descriptive characteristics • Identifying common themes or checking for specific topics in responses to open-ended survey questions • Tallying numbers of documents in categories based on their content • Classifying open-ended feedback into topics • Cross-checking the results of manual synthesis of responses to open-ended survey questions
Summarisation	Automatically create a readable summary of the key points in a block of text	• Filtering of documents or to separate relevant and irrelevant material • Extracting key points from interview transcripts or survey responses for further analysis • Cross-checking manual summaries and synthesis of literature

Acknowledgements

Adrian Field and Julian King provided very helpful feedback on a draft.
Cover photo by Wesley Tingey on Unsplash.

Contents

Motivation

Evaluators and researchers often work with text-based data such as open-ended survey questions, interview transcripts, organisational documents, and other written reports. Text data requires interpretation and analysis before it can be used, but this can be time-consuming or infeasible to do manually for large volumes of text. In some contexts, automated text analysis using freely available tools can help to streamline and augment the process of manual analysis. Automated text analysis combined with manual review can improve the efficiency and quality of evaluation and research work by providing a new approach to analysing text data and as a cross-check for human analysis.

Early approaches to automated text analysis were based on identifying keywords or phrases in text. This is easy to do but may not identify complex topics or concepts and relies on the researcher choosing a suitable set of keywords. Natural language processing (NLP) models are applications of machine learning techniques that have become advanced enough to support real-world research and evaluation tasks. NLP models move beyond counting keywords and embody a degree of semantic 'understanding' of language, derived from training a statistical model to recognise linguistic patterns in large volumes of digitised text. This enables NLP models to detect complex concepts in text that may be expressed in different ways.

Automated text analysis cannot yet fully replace human review and synthesis of text data, but it can assist evaluators and other researchers by:

- Extracting topics, sentiments, and themes from text to guide and focus manual analysis
- Acting as a ‘sense check’ for manual analysis and potentially picking up information that may be missed in manual review
- Augmenting analysis of quantitative data by extracting characteristics or other information about data subjects from associated qualitative data
- Providing a ‘first pass’ filter in large literature or document reviews to help identify the most relevant material

This paper aims to give a brief introduction to some automated text analysis techniques and how these can be used in real-world research and evaluation work. The focus is on what is possible rather than technical details but brief summaries of how the methods work are included. Three applications of text analysis that may be most useful for researchers and evaluators are covered: sentiment analysis, topic detection, and summarisation. Each application is illustrated in the context of a real-world task, and the data required to make it work is described.[1] Limitations of and caveats for using NLP models are also discussed so that evaluators and researchers can gain a realistic understanding of how automated text analysis can be used appropriately and effectively.

[1] All examples are based on actual research and evaluation projects, but details have been changed to protect the privacy of survey respondents and other confidential information.

Sentiment analysis

Sentiment analysis aims to classify the sentiment expressed in a short block of text as 'positive', 'neutral', or 'negative'. More generally, a 'sentiment score' can be assigned to a block of text, measuring sentiment along a continuous scale from very negative to very positive, represented by a number between 0 and 100. The overall sentiment of a block of text can be assessed, or sentiments associated with separate topics within the text can be identified (combining sentiment analysis with automated topic detection is discussed below).

Uses in research and evaluation

Sentiment analysis can be applied in various ways, including:

- Classifying and ranking responses to open-ended survey questions, such as questions about experiences with a product or service
- Classifying interview responses to specific questions about experience or satisfaction
- Processing customer or client feedback or reviews such as those collected via feedback forms on websites or post-experience surveys
- Summarising sentiments expressed in social media posts by groups of users or about certain topics of interest

Example: Classifying sentiments from survey responses

People who were assisted by a local public health service were surveyed about their experiences and asked: *If you had to describe to a friend or whānau member in two or three words or short sentences what the support you received was like, what would you say?*

Example responses and automated classification of the sentiment associated with these are shown in the table below. Results from a sentiment analysis model applied to each response are shown in terms of the probability of negative, neutral, or positive sentiment.[2] These probabilities can be analysed and summarised directly, or each response can be classified into a primary category based on the sentiment with highest probability.

	Sentiment classification (%)		
Survey response	**Negative**	**Neutral**	**Positive**
It was nice to be treated like a person, not a number.	0.2	6.8	93.0
It was reasonable but nothing earth shattering. I did appreciate the help with childcare.	0.4	28.6	71.0
I wouldn't use this service again as I ended up having to pay for testing that I thought was supposed to be free.	97.5	2.2	0.3
A wonderful and caring team; I was very touched by this service.	0.2	0.6	99.2
Responses to my requests were a little slow but overall it was fine.	2.4	92.7	4.8

[2] Sentiment analysis was done using the bertweet-base-sentiment-analysis model (see below for a description of this model).

Data requirements for sentiment analysis

Sentiment analysis is best applied to short blocks of text, from one sentence to a short paragraph. This text should be collected in a way that encourages unambiguous expression of positive or negative sentiment. Survey and interview questions that will be used for sentiment analysis should be carefully designed to avoid respondents to make contrasts and comparisons, and to encourage respondents to answer directly.

For example, contrasting statements like “the food was terrible, but the conference presenters were excellent!” may be misinterpreted by automated sentiment analysis. Where possible, open-ended questions should be designed to encourage respondents to give opinions about only one topic or aspect at a time. Another option is to combine sentiment analysis and topic detection (see below) to extract separate sentiments for each topic where these are combined.

Overview of how sentiment analysis works

A statistical (machine learning) model is first trained to recognise different types of sentiment in text based on millions of examples of statements that have already been classified into positive, neutral, or negative sentiment categories. This is done by first processing raw text into numerical data using ‘word embeddings’: semantically similar words are identified and grouped, then given similar numerical values that can be easily manipulated by a computer.

This numerical data is fed into a model that aims to predict the sentiment of the associated text. Across the dataset, the model is trained to recognise relationships between numerical word embeddings and positive, neutral, or negative sentiment capture in the training data. This is done by adjusting parameters of the model until it can predict the sentiments in the training data as accurately as possible.

Sentiment analysis is then performed by giving the trained model new text to be analysed. This is also first processed into numerical form in the same way as the training data. The trained model generates a prediction of the associated sentiment for each numerical representation of the data being analysed. This can be expressed as a numerical sentiment score along a continuous range, and/or categorised as positive, neutral, or negative depending on the value of the score. An example of a freely available model for sentiment analysis is bertweet-base-sentiment-analysis, which was trained on text from 850 million English-language tweets from Twitter and further refined on 40,000 tweets classified by sentiment. Models are also available for other languages.

Topic detection

Topic detection aims to identify the topics or concepts that are referred to in a block of text. This can be done in two different ways, depending on whether potential topics of interest are known in advance or not:

- *Topic classification:* Assesses whether individual blocks of text relate to one or more pre-specified topics. This involves estimating a probability between 0 and 1 that a block of text relates to a given topic, given the content of that text. Topic classification is done using a pre-trained language model and can be used to check for topics in any number of blocks of text.
- *Topic modelling:* Identifies the common topic(s) in many blocks of text. Topics are expressed as lists of keywords that appear to be related within the blocks of text being organised. These lists of keywords are then classified manually into discrete topics. Topic modelling is based on statistical analysis of text and requires a reasonable number of blocks of text to be effective (typically in the hundreds). This does not require a pre-trained language model, but the results may need some manual work to refine and interpret the identified topics.

Machine learning approaches to topic classification do not simply look for the presence of topic keywords in text. Instead, language models are used to assess whether the text is related to the topic even if it doesn't include those specific words, based on a model of ways that topics can be expressed in words.

Uses of topic detection in research evaluation

Real-world research and evaluation tasks where topic detection can be applied include:

- Classifying data subjects into groups based on topics included in descriptive characteristics, to facilitate further analysis of subjects across those characteristic groups.
- Identifying and/or categorising common themes in responses to open-ended survey questions.
- Checking for the presence of specific topics of interest in open-ended survey responses.
- Tallying numbers of documents in categories based on their content.
- Classifying open-ended feedback provided by participants in a programme into topics.
- Quickly cross-checking the results of manual synthesis of responses to open-ended survey questions to ensure that no important topics were missed.

Example of topic classification: Classifying data subjects based on descriptions

Descriptions of projects to invest in regional facilities and other activities funded by the NZ government's Provincial Growth Fund (PGF) were analysed to determine if they may be related to *employment*, *youth*, or *tourism*, as these topics were of particular interest.[3] The probabilities associated with these topics for example project descriptions are shown in the table below.[4] These

[3] Raw data on PGF projects is available at https://www.growregions.govt.nz/established-funds/what-we-have-funded/.

[4] Topic classification was done using the bart-large-mnli model. See below for a description of this model.

probabilities can be analysed directly, or projects can be assigned to one or more topics detected with sufficiently high probability.

	Topic probabilities (%)		
Project description	**Employment**	**Youth**	**Tourism**
To deliver a food production hub in Ruatahuna specialising in honey initially, PGF funding sought to support infrastructure requirements of the hub including fitting out the production and packing facility	43.6	4.8	0.0
This programme will see an existing tourist accommodation facility into a Life Skills Youth training facility for 60 NEETS who will be provided a pathway and pastoral care to navigate the transition from leaving school into the workforce, or into further training. The programme has a collaborative approach with iwi, hapu, the community and other service providers within the immediate and wider region.	99.5	96.9	93.7
Developing a hot pools and spa complex in Methven, Canterbury that uses solar energy as the primary means of water heating.	33.6	10.8	9.4
To assist with the relocation of a public weighbridge facility from between the entertainment precinct and mixed use commercial area of Ahuriri, Napier to the Pandora industrial zone, Napier	63.8	1.7	0.1
Funding provided for the enhancement & remediation of the Farm including fencing, fertiliser, water troughs, and cattle yard repair.	3.1	9.8	0.0

Example of topic modelling: Identifying common topics in open-ended survey responses

Text descriptions of all projects funded by the PGF were modelled to find common topics referenced in these descriptions.[5] Four topics detected within these descriptions and the top five keywords associated with each topic are shown below. From these lists of keywords, descriptive topics can be created. Each project can also be assigned to one or more topics, based on the existence of keywords contained in its description.

Topic	1	2	3	4
Keyword 1	marae	programme	road	water
Keyword 2	cluster	employment	toilet	fencing
Keyword 3	renovation	training	facilities	upgrade
Keyword 4	refurbishment	people	develop	maintenance
Keyword 5	renovations	rangitahi	Improve	existing

[5] Latent Dirichlet Allocation (LDA) was used to model topics in this data. See below for a description of LDA.

Data requirements for topic detection

The key requirement is blocks of text to be analysed. These may range from one sentence to multiple paragraphs. Topic classification can be done on one or more blocks of text and requires a list of pre-specified topics expressed as a keyword or short phrase. For complex topics, it may be useful to specify several alternative keywords or phrases and combine the results from these.

Topic modelling typically requires hundreds of blocks of text to detect distinct topics, but the topics do not need to be pre-specified. Prior to topic modelling, it is usually necessary to remove 'stop words' such as *the*, *a*, *and*, *but*, etc from the text, correct spelling errors, and remove any other frequently occurring text that does not relate to topics, such as numbers. These data cleaning tasks can be done in a semi-automated way using topic modelling software.

Overview of how topic detection works

Topic classification

Like sentiment analysis, a topic classification model can be trained using large amounts of text that has already been tagged to topics. This model learns the association (or lack of association) between numerical representations of text and topics. A trained model can then be used to assign new text to topics based on the modelled relationships. This approach works best for specialised topics (e.g., technical, or subject-specific topics) where a large training dataset is available.

Alternatively, a natural language inference (NLI) model can be trained on a large set of pairs of text 'premises' and 'hypotheses' that are either true or false. For example, a premise of "a soccer game with multiple males playing" and a hypothesis of "some men are playing a sport", which is true. This can then be used to estimate whether a new block of text relates to a given topic by posing the text as a premise and the topic as a hypothesis. With this setup, a NLI model can estimate the probability that the hypothesis is true, and hence that the text relates to the given topic. An example of this approach is the bart-large-mnli model which was originally developed by Facebook and further trained on 433,000 premise/hypothesis sentence pairs.

Topic modelling

Extracting topics from text where the topics are not known in advance is typically done by constructing a statistical model of topics contained in blocks of text based on the words and phrases that tend to occur together. There are several methods for doing this, with the most common method known as Latent Dirichlet Allocation (LDA). Unlike sentiment analysis and topic classification, this is a purely statistical approach that does not require a pre-trained language model, although trained models can produce better results.

The output of topic modelling using a method such as LDA is multiple lists of keywords that appear to be linked via a common topic across the blocks of text that were analysed. Human input is usually required to construct representative topics from the output lists. In some cases, it may also be necessary to combine similar topics detected by the topic model so that the final topics are sufficiently differentiated. A limitation of topic modelling is that the number of distinct topics to be detected needs to be specified, but this is not usually known in advance. Some trial and error is often necessary to get sensible results by changing the number of topics.

Summarisation

NLP models can generate automated human-readable summaries of the key points in blocks of text. This is not a substitute for human review and synthesis, particularly for specialised or technical topics. However, automated summarisation can be used to augment manual review of literature and other documents, to speed up and improve the quality of the process.

Uses in research and evaluation

Automated summarisation can be used effectively for:

- 'First pass' filtering of documents or literature to separate relevant and irrelevant material
- Extracting key points from interview transcripts or survey responses for further analysis
- Cross-checking manual summaries to ensure that key topics were not missed

Example: Rapidly processing one-page summaries of documents

The blocks of text in the table below were taken from the summary section of a World Bank evaluation report.[6] Automated summaries of around 50 words were generated using the bart-large-cnn model (see below).

Original text	Automated summary
Municipal solid waste—waste generated mainly from residential and commercial sources—has emerged as one of the most pressing challenges across the world, with growing public health, environmental, social, and economic costs. By 2050, fast-growing large- and medium-size cities will nearly double the waste generation in lower-middle-income countries and upper-middle- income countries. Low-income countries (LICs), where most waste is disposed of in open dumps, are on a trajectory to triple their municipal solid waste generation by 2050. Historically, the causes and effects of municipal solid waste were considered local or regional. However, with increasing volumes and changing waste composition, municipal solid waste has become a global challenge. *107 words*	By 2050, fast-growing large- and medium-size cities will nearly double the waste generation in lower-middle-income countries. Low- Income countries (LICs), where most waste is disposed of in open dumps, are on a trajectory to triple their municipal solid waste generation by 2050. Growing public health, environmental, social, and economic costs. *50 words*
The waste hierarchy and the circular economy are sustainable alternatives to the traditional linear (take-make-dispose) economic model. The traditional economic model approaches the waste value chain as a linear sequence in which resources are extracted from the environment (take), manufactured into goods (make), and discarded when they are no longer needed or wanted (dispose). The waste hierarchy approach lays out a more nuanced but still linear set of disposal options and establishes a ranking among them from most to least preferable. Waste prevention and reuse are the most preferred options, followed by recycling, then recovery (for example, composting and waste to energy); waste disposal through landfills should be the very last resort. The circular economy approach closes the loop in relation to extraction, manufacturing, and disposal by advocating for designing products to reduce waste, using products and materials for as long as possible, and recycling materials from end-of-life products back into the economy. *153 words*	The traditional economic model approaches the waste value chain as a linear sequence in which resources are extracted from the environment. The waste hierarchy approach lays out a more nuanced but still linear set of disposal options. The circular economy approach closes the loop in relation to extraction, manufacturing, and disposal. *51 words*

Data requirements for summarisation

Summarisation is applied directly to one or more text documents to be summarised. These can range from one paragraph to several pages. For longer documents, better results may be achieved by

[6] *Transitioning to a Circular Economy: An evaluation of the World Bank Group's support for municipal solid waste management (2010-20)*.

applying summarisation to existing executive summaries or abstracts, if available. 'Off the shelf' summarisation models currently available tend to have a limit of around 1,000 words for the original text, so longer pieces of text will need to be separated into chunks for summarising. A simple approach is to apply summarisation to each paragraph separately.

Overview of how summarisation works

There are two main types of methods for generating automated summaries from existing text:

- *Extractive summarisation*: This creates a summary by extracting key sentences from the original text. These are selected based on statistical measures of the 'importance' of each sentence in the overall text. For example, a sentence that contains words that often occur elsewhere in the text may have high importance. The length of the summary can be controlled by selecting the number of sentences to be extracted from the original, or a cut-off for the degree of 'importance' for including a sentence in the summary.
- *Abstractive summarisation*: A summary is constructed by identifying key topics and passages in a text. Summaries generated in this way can be further improved by training a language model on pairs of text and summaries written by humans, so that it learns the structure of a good summary. An example of this approach is the bart-large-cnn model, which was fine-tuned on a dataset of 300,000 news articles from CNN and the Daily Mail with accompanying human-written summaries.

In both cases, it is possible to adjust the length of the summaries that are generated. This is usually done by specifying a target range for the length of the summary (e.g., 50 to 100 words).

Combining methods

Topic detection and sentiment analysis can be combined to first detect topics in text and then analyse the sentiment associated with these. This is known as aspect-based sentiment analysis, and the distribution of positive, neutral, and negative sentiment associated with each topic (or 'aspect') can be summarised. The topics for which sentiment is to be analysed can be pre-specified by the researcher if there are known topics of interest (topic classification) or detected automatically otherwise (topic modelling).

This may be useful for analysing open-ended survey or interview responses that may include a range of topics. Topics identified in the responses can be categorised according to the sentiment associated with them. With a large enough number of responses, it is also possible to analyse if different groups of respondents expressed different sentiments associated with certain topics, which may suggest useful issues to explore in follow-up or further research.

Preparing text data for automated analysis

General data cleaning

While NLP models have become very sophisticated, the results of automated text analysis can be improved if data is suitably collected and processed prior to analysis. The following steps should be performed prior to automated analysis:

- Open-ended survey or interview questions should be expressed in ways that encourage clear and simple responses, avoiding technical jargon, slang, and idioms where possible.
- A spelling checker should be used to correct common spelling errors.
- Expand acronyms and abbreviations where possible, particularly if these are non-standard or specialised terms that are not commonly used.
- Prior to automated summarisation, remove extraneous text, e.g., copyright notices, disclaimers, author biographies, etc.

Considerations for Te Reo Māori

In Aotearoa New Zealand it is common for Te Reo Māori words to be mixed with English. As most 'off the shelf' language models were trained using international datasets, these models will have no or very limited ability to correctly recognise topics or sentiments expressed using a combination of Te Reo Māori and English. Prior to using automated text analysis on such data, consideration should be given to translating Te Reo Māori words that have a very similar English equivalent in when used certain contexts, such as whānau, motu, mahi, and tamariki. Translation may be less appropriate for more complex concepts (e.g., tikanga, manākitanga, and kaitiakitanga), or if a significant proportion of Te Reo Māori is used. At present, Māori language models are not widely available for automated text analysis but are under active development.[7]

[7] See for example the Papa Reo language platform being developed by Te Hiku Media.

Validation

It is important to validate the results of automated text analysis to have confidence in its accuracy. This can be done simply by taking a random sample of original text and comparing the results of manual analysis to the results of automated analysis. The manual assessment should be done in a 'blind' way, i.e., without knowing the results of the automated analysis for any given block of text, so that the human reviewer is not influenced by the automated results.

The proportion of the sample for which the manual and automated results agree can be calculated as a measure of accuracy of the automated analysis. For example, the proportion of sentiments in the sample classified as positive, neutral, or negative by both a manual reviewer and automated analysis can be calculated. It is very likely that there will be some differences due to the limitations of NLP models and the somewhat subjective nature of many text analysis tasks. However, any substantial differences can be investigated to determine if the automated results are unreliable.

Limitations of automated analysis

In the fields of research and evaluation, text data is often the source of key insights and supports evaluative conclusions, which means it needs to be reliable and robust. While NLP models have become very sophisticated, these are best used in conjunction with human analysis rather than as a replacement for it. Results from 'human in the loop' text analysis are typically superior to fully automated text analysis. For example, automated topic modelling may detect two separate but similar topics that a human would choose to combine. At a minimum, a sample of the results of automated analysis should be checked manually for accuracy and bias.

The following other limitations should be considered by researchers and evaluators who wish to use automated text analysis in their work.

Biases in training data may be reflected in the results of automated analysis

A common concern when using machine learning or 'artificial intelligence' (AI) technologies is the potential for biased outputs, which may be very problematic in some contexts. Such biases originate from the data used to train machine learning models, rather than from the models themselves. A well-known example is the problem of gender bias associated with occupations. Due to historical and social stereotypes being incorporated in training data, machine learning models may be more likely to recognise some occupations as 'male' and others as 'female'. Similar biases may arise with other potentially sensitive characteristics such as ethnicity or sexual orientation.

Bias risks can be mitigated by avoiding inappropriate uses of automated text analysis. For example, automated text analysis should not be used to try to infer personal characteristics of survey respondents from answers to questions that do not directly ask about these characteristics. It is also risky to disaggregate the results of automated text analysis by characteristics that may be subject to bias. For example, people with different ethnic backgrounds may tend to express positive or negative sentiments in different ways and disaggregating the results of sentiment analysis by ethnicity may give misleading conclusions.

In general, it is important to remember that 'off the shelf' language models have been trained using specific datasets and the output of such models may reflect any existing biases in those data. Where possible, verify the training data that was used for a model, and search for any reports of bias within that data.

NLP models do not cope well with some aspects of human language

Despite recent advances, NLP models may not be able to accurately analyse certain aspects of human text or speech such as idioms, irony, or sarcasm. Open-ended survey questions can be structured in a way that discourages these, but interview transcripts may be more problematic, particularly if conducted informally.

Specialised technical language or jargon may also create challenges, as this is less likely to have featured in the training data used by 'off the shelf' language models. Accurate analysis of such text may require a customised language model, which may not be feasible to develop for a one-off evaluation or research project. However, it may be possible replace jargon with plain language prior to automated analysis using a simple 'find and replace' approach for relevant keywords.

"Black box" models may not be well accepted

While the basics of automated text analysis are relatively easy to describe, most NLP models are 'black boxes' from the perspective of most evaluators or researchers and their clients. In many cases the source code and training data of the models is open and can be inspected but understanding these requires specialist knowledge and may be time-consuming. In addition, the mechanics of NLP models use sophisticated mathematical and statistical concepts that are hard to explain to people who do not have the necessary background knowledge.

As a counterpoint, it can also be argued that human analysis of text data is equally a 'black box', and it is very difficult to be transparent about how a human researcher reached judgements about sentiments or topics contained in text. Trust in automated text analysis can be increased by manually reviewing a sample of the results and comparing the findings, to demonstrate the overall accuracy of automated analysis.

Automated analysis can have significant setup costs

The setup costs of automated analysis include testing and fine-tuning models, cleaning and preparing text data prior to analysis (as described above), and manually reviewing a sample of the results for accuracy. Typically, this involves an iterative process where the automated analysis is improved based on manual review, as fully automated analysis is generally inferior to 'human in the loop' analysis. Automate analysis can only be justified if the amount of text to be analysed is large enough so that the setup costs are outweighed by the overall time savings and/or improved quality of the analysis.

About Schiff Consulting

Schiff Consulting applies economics, data science, and evaluation to help public and private sector clients in New Zealand and abroad learn, change, and solve problems. Its principal, Dr Aaron Schiff, has over 20 years of experience in economic analysis and modelling, cost-benefit analysis, quantitative outcome and impact evaluation, statistical analysis, and forecasting. Schiff Consulting has served leading organisations across many sectors including transport, health, local government, energy, infrastructure, telecommunications, and social enterprise. Schiff Consulting also regularly collaborates with other leading consultancies.

www.schiff.nz
+64 9 336 1323
Auckland, New Zealand