Analysis of Open Answers to Survey Questions through Interactive Clustering and Topic Extraction

Gavagai, Stockholm, Sweden

September 1, 2020

Abstract

This whitepaper describes design principles for and the implementation of The Gavagai Explorer—an application which builds on interactive text clustering to extract topics from topically coherent text sets such as open text answers to surveys or questionnaires.

An automated system is quick, consistent, and has full coverage over the study material. A system allows an analyst to analyze more answers in a given time period; provides the same initial results regardless of who does the analysis, reducing the risks of inter-rater discrepancy; and does not miss responses due to fatigue or boredom. These factors reduce the cost and increase the reliability of the service. The most important feature, however, is relieving the human analyst from the frustrating aspects of the coding task, freeing the effort to the central challenge of understanding salient topics as expressed in the text.

Gavagai Explorer is available on-line at http://gavagai.io

Open Answers to Surveys are a Challenge

Open answers in surveys and questionnaires are a challenge for analysts: how to report the collected responses together with more quantitative data elicited from respondents is not obvious. Typically, a human analyst, or a team of analysts, have been given the text responses in question, with an editorially determined coding scheme, a kind of manual annotation that references when certain phrases are used to be classified under specific

categories, discussed and revised at intervals. The task of the analysts is to label the responses according to the coding scheme and to extract samples from the responses to anchor the labels in the data.

This coding or manual annotation procedure, converting the open responses into a structured form, requires time and expertise on the part of the analyst, both of which come at a cost. The effort involved in coding open answers is simultaneously intellectually non-trivial and demanding, but still monotonous: analyst fatigue and frustration risks leading to both between-analyst and within analyst inconsistencies over time in reporting. This challenge is well-established both in the market research field and in scientific studies.[1]

The task is related to text categorisation, but not identical to it: no pre-defined palette of categories is available, and the texts are by definition topically aligned. The differences in most customer cases have to do not with topic per se, but with polarity or attitude vis-a-vis the topic of the question and in what facet of that given topic motivates the attitude expressed in the answer. In example (1) some extracts from reviews for a hotel are shown.

(1)

A. I would definitely recommend this hotel, the location was great!
B. Had I known, I would NOT have chosen this hotel for my busy work visit in which I needed quiet time in the hotel to do work.
C. Modern, stylish hotel with numerous, pretty decent restaurants in the area!

Human information analysts take about 1 minute to categorise an abstract[2], and this is for a case where the categories are already given. If the task is to explore a set of responses and define and revise the categories or labels to indicate the topical content of the texts, this involves more effort and is likely to require more time per item.

[1] E.g. O'Cathain and Thomas (2004) and many others.

[2] E.g. McCallum et al. (1999); Schohn and Cohn (2000); Macskassy et al. (1998) and many others.

This paper describes a productivity tool for interactive coding, i.e. exploring and assigning thematic labels to open responses, based on a back-end technology which learns terminology and semantic relations from text.[3]

Use Case of The Gavagai Explorer

The purpose of including open questions in a survey is to explore the underlying motivations of the respondents with respect to some topic of interest. These motivations can be known in advance, they may be somewhat predictable, or they may be entirely unknown to the analyst. The resulting analysis, which is intended to give insights form the basis of e.g. strategic market decisions, product improvements, management strategies or other actions for the analyst, will be a set of such topics. These topics with relevant quotes extracted from responses, reported together with their relative strengths and quantitative statistics such as Topic frequency or Topic Sentiment on the numbers of respondents involved in discussing each topic create a picture for the analyst to start to understand the text data. If a relevant structured categorisation scheme is available, e.g. an ontology, a gazetteer or a knowledge graph of some sort, the topics should be related to those categories as well; if a relevant sentiment analysis palette is available[4], this can be brought in to quantify the attitude of the respondents with respect to each topic.

The Nature of Human Language

The ambiguity, vagueness, and fluidity over time of human vocabulary is often described as a problem for modelling human communication. This perspective does not do justice to the nature of human communication. The adaptability of human vocabulary and thus the entire human communication system is in fact useful: it allows new terms to be coined,

[3] This approach builds on a long-standing strand of research in information retrieval which builds on interplay between similarity based clustering and end-user assessment of clusters, such as Jardine and van Rijsbergen (1971); Cutting et al. (1992); Pirolli et al. (1996); Sanderson and Croft (1999) and many others.

[4] Karlgren et al. (2012)

established terms to be recruited into service ad hoc to fit the needs of some discourse, and various discourses to be associated or contrasted through term choice. The challenge for the analyst of our specific use case is in fact exactly the reason why open answers are useful. If the choice of words were entirely predictable, the information captured through open answers would be so much less rich and valuable for the analyst.

(2)

A. The appearance of the text, the quality of its design and polish.
B. How enjoyable and fun it is, how it addresses its readers, and who has written it.
C. Who wrote it and why.
D. Does it speak to me?

The examples in (2) are translated to English from a survey on how respondents assess the trustworthiness of textual materials referring to various qualities of the texts the respondents have read.[4] There are at least two topics in these four responses: the *source* of the text and the *audience design* of the text. The first topic was an expected topic, the second somewhat unexpected, and it would have been difficult for an editor to instruct a coding scheme to make note of terms such as *speak*, and *address* before the fact.

This sort of information is exactly what the study was designed to find. The intention underlying the design presented here is to empower the analyst to fold together *X*s and *Y*s into a topically coherent topic, retaining the variation found in the material leaving the underlying data unchanged, not to normalise the behaviour of the respondents into a uniform vocabulary given before the responses and lose out on the rich variation of data.

[4] The survey was performed in the Fall of 2016 to explore the attitudes to digital tools in teaching among students. http://www.berattarministeriet.se/undersokning/

Design Principles of The Gavagai Explorer:

Design Principle 1: Empowering Analysts, Not Replacing Them

A repetitive and frustrating task often is understood as a candidate for full automation. Our design is instead based on the work practice of human analysts, and intended to afford a human analyst tools to work with the text smoothly and painlessly, leaving the human effort to be expended on the most crucial and demanding task of content analysis, but freeing the analyst from keeping track of consistency.

Design principle 2: Incremental refinement in clustering pipeline

The assumption of interaction designers is often that users are best served by automation. Our design is a departure from that assumption based on the idea that no system could know exactly what the analyst will want given any data set. We want our system to go beyond a one-shot dialog. The dialog builds on incremental specialization of the analysis: in a few iterations of the data set, the analyst can achieve a stable clustering to save and report.

Design principle 3: Errors do not matter

The assumptions made by the system, however well its algorithms are designed and however well established its background knowledge is, are often daring and sometimes mistaken. The design is intended to display analyses, and to allow the analyst to correct misclusterings with little effort, with a high degree of interactivity. The above principle of incremental refinement alleviates the presence of errors — the analyst is able to find topics in the texts, even if some of the first clusters were irrelevant or overlapping.

Design principle 4: Representation in surface terms

The end result of the analysis is a knowledge representation through which the set of texts can be understood better. This structure can be saved for future incoming data sets, e.g. a before-and-after study or a periodically

repeated survey over some population. We want the knowledge representation to be inspectable, reportable, and editable by a human analyst without specialist knowledge. The representation is entirely in surface terms, for that purpose.

Design principle 5: No dependence on outside resources

We want the system to be portable to various languages, various domains of application, and various cultural areas. We do not want it to rely on costly or cumbersome lexical or encyclopedic resources which may not be available in all languages. The system is designed not to need anything from the analyst but the texts under consideration. This also means the system can be deployed on premise or on other cloud based solutions if needed.

Implementation

The functionality on which the system is built automatically clusters the documents into Topics by frequency count. This creates clusters of documents that share topically important terms. In other words you can gain an understanding of how often Topics appear in the dataset and skip the manual analysis of reading and annotating manually how often Topics appear. In addition, you gain insights about sentimental adjectives on a Topic basis, and the system will be able to give you the related Topics for each Topic. Following the above design principles, these clusters are then displayed to the analyst for consideration.

The main actions for the analyst are:

- joining existing closely related Topics for example, *Staff* and *The Front Desk*
- discarding Topics that are of no interest for example, *Hotel*
- working on what terms characterise a Topic by approving synonyms suggested by the system or entering them manually for example, adding the term *Penthouse Suite* to the *Topic: Room*

Quantification of Qualitative Data

By using text analytics, the analyst should be able to use statistical analytic methods to derive insights from what the text was actually about. This is completely opposed to "Word Clouds" or similar methods where the results are ambiguous. The analyst should be able to quantify topics and draw robust conclusions from the results of the analysis. The Explorer is able to return multiple numerical data points from text analytics:

- Topic Frequency: How often topics are mentioned in the entire text. This would be similar to manually coding or annotating text data and then counting it up.
- Topic Sentiment: This is how often certain Topics have Sentimental adjectives describing a topic. The Gavagai Explorer supports positivity, negativity and five other sentiments
- Related Topics: How often certain Topics appear as a related Topic in the text data.

Text clustering

Lexical clustering builds on measures of term specificity to select which terms to use as clustering features, which requires general language data to be able to assess how specific or general a term is. Clustering by terms is fairly sensitive to genre-specific and topical usage, since a term which has high specificity in general language may have little utility in the context being examined.

Most standard lexically based clustering algorithms give similar results; we use a clustering algorithm based on insights from our previous research results on distributional semantics,[5] and we find that improving response speed and capacity of the system are more important to address (given

[5] Gyllensten and Sahlgren (2015)

Design principles 2 and 3 above) than marginal improvements in cluster quality.[6]

The example sentences from a hotel review data set given in (1) were all in the first iteration clustered together under the label *hotel*. They should in most scenarios not end up being clustered on *hotel* but instead on *location* (for samples (1-a) and (1-c)) and *work* (for sample (1-b)) instead.

A term such as *hotel* in hotel reviews is not a useful clustering feature. Addressing this challenge requires automatically reweighting term specificity during the clustering process, and, most importantly, as our system currently does, consulting the analyst to see if the clustering terms are appropriate and informative.

Synonyms

The nature of human language being as it is, we can expect many answers to diverge from the expected terminology. There will be many ways to say the same thing and you want them all in the same Topic after the analysis

[6] This is in keeping with earlier results comparing different text clustering systems, comparing their output with human assessments. There are differences, but they are comparatively small. Roussinov and Chen (1999)

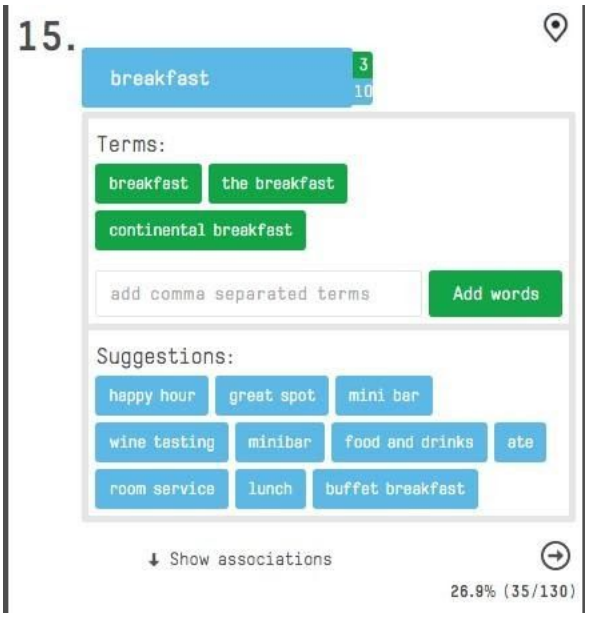

Figure 1: The system suggests synonyms for *breakfast*, including multi-word terms.

process. The topic is represented by a set of terms which are prevalent in the texts clustered into that topic, and using a lexicon learned from text[7] in the target language, the system suggests synonyms from the uploaded text data to increase the coverage of that topic such as the term *food and drinks* for Topic: *breakfast* as shown in the example in Figure 1. The analyst is also able to freely enter terms to enrich the representation of a topic.

Multi-word terms

Most written languages build on white-space separated words, which is very convenient for tokenization of the input stream in text processing. Many languages — and English is especially liberal in this respect — formulate multi-word compound terms quite freely, and all languages have set phrases such as *kick the bucket* and some degree of lexicalised multi-word terms, not least names such as *San Francisco* but also technical

[7] Sahlgren et al. (2016)

terms such as *linear accelerator* or *bed linen*. A system built for lexical clustering needs to note these multi word expressions or n-grams as they occur and include them as a clustering basis. Our system is built to identify n-grams incrementally[8], as they appear in streaming data, and use this to propose multi-word terms found in the text.

Manipulating clusters

The action of *joining* clusters into one common topic is a frequent operation to refine the end result, and Gavagai Explorer supports joining through simple direct manipulation as shown in Figure 2. Similarly, clusters of low utility can be *discarded*, and the items constituting it are redistributed over other clusters instead. An example is shown in Figure 3. In this way, the content of the clusters are iteratively refined with simple and reversible point-and-click manipulation.

Figure 2: Joining clusters: the cluster labeled *service* is a candidate for being merged with *offered*.

[8] Sahlgren et al. (2016)

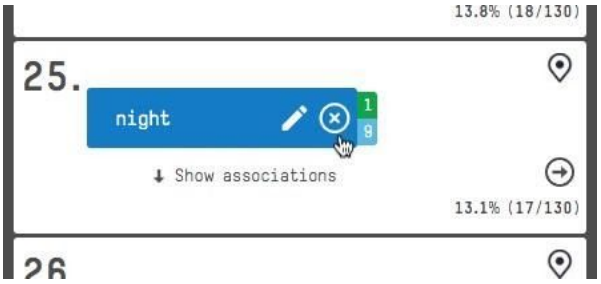

Figure 3: The cluster labeled *night* may not be informative for hotel reviews.
such as a collection of newsprint, or a Wikipedia snapshot), and it still requires the analyst to be handy in the source language of the texts.

Saving and Reusing a Template Model

A coding scheme that is, how Topics are chosen to be formed with certain Terms, after being used for a data set can be saved and reused for future incoming data sets. For example, when you are comparing two surveys from different time periods or different surveys from two geographic locations then using the template models allows you to compare Topics across the surveys.

Handling Several Languages

Analysis of responses must as a rule be done in the language the responses were submitted. Translation will always introduce an interpretation of the data in play. Gavagai Explorer is built to be language agnostic and handles any human language. To deliver reliable high-quality synonyms the system needs to have had access to some collection of general texts in the source language.

When there are several similar data sets in multiple languages, for example a survey done in several different language markets, The Explorer can utilize the template model function and apply a model consistently across the languages. Thereby, allowing an analyst to compare Topics across markets and glean insights from text data from multiple languages.

Without translating the underlying data or modifying the source data in any way.

Preserving the Original Data to be Further Analyzed

After analyzing the Topics on a verbatim or row by row basis, The Explorer ensures other columns are taken into account by preserving the original data set. Thus, if other columns such as NPS, grade, demographics or interesting metadata columns are present then the analyst can begin to slice the Topic results and analysis by these metadata columns for deepening possible insights. For example, Topic analysis by age cohort to see which Topics are deemed most important to which age cohort or Topic analysis by NPS category.

Visualizing all the Results

After the Explorer extracts the Topic results, the analyst can visualize the results by entering the Dashboard and can share the results of the analysis with a link. They can also create new custom graphs by using other metadata columns to slice the Topics, Sentiment, and or Grade Scores by different demographics. Thus, getting a better understanding of the actual text data..

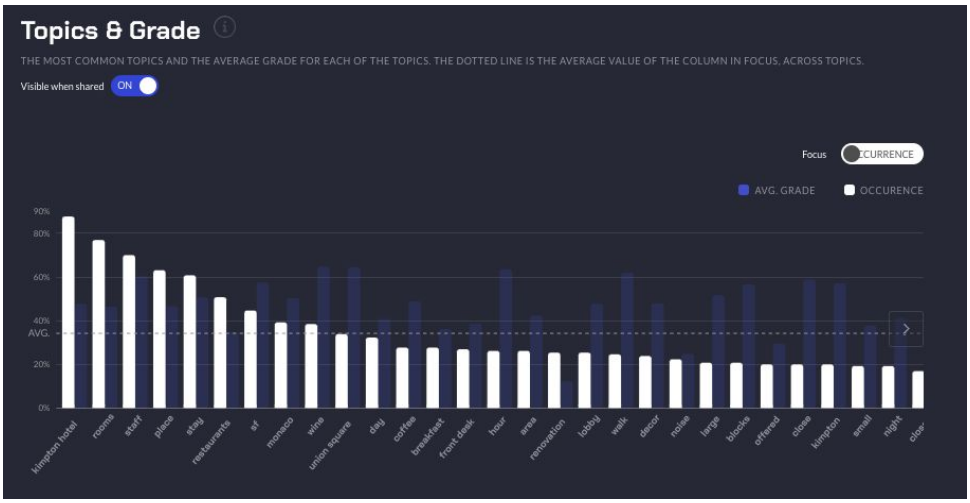

Figure 4: Visual representation of Topics in a Bar chart with the height of the white bars representing how often people are mentioning a Topic in percentage from the entire survey. The blue bars represent the correlated Grade score for all the reviews mentioning that Topic.

Case studies

We present here short abstracts of case studies where the Explorer as described above has been used. They serve to illustrate the versatility of the tool, in application to multi-lingual and multi-cultural data, very open questions of wide-ranging topics, and drilling down into subtopics of customer reviews.

Attitude towards gender equality in seven cultural areas

In 2016, Gavagai conducted a Gender Equality Study in Saudi Arabia, United Arab Emirates, Russia, Sweden, Colombia, Mexico and Brazil commissioned

by the Swedish Institute[9], a Swedish government agency with the task of creating goodwill for Sweden through public diplomacy efforts. This survey is part of an effort to monitor awareness of some aspects of Swedish society in focus of Swedish foreign policy. The study collected 9800 free-text answers to open-ended survey questions through survey partners in the various cultural areas, and the answers and the results were analyzed with Gavagai Explorer.[10]

The findings show interesting differences in the way the various cultural areas approach the notion of gender equality. Employing domestics to redress home chore imbalances, the attitude to the label ”feminism”, the view on who in the family unit should be involved in important decisions on e.g. economy all vary across the cultural areas in interesting ways: how those differences can be understood and explained was made possible by identifying topical topics among the items with attitudinal loading. As one example we found a clear difference across cultural areas with respect to ”feminism”. The question as given in (4) gave attitudinal results as shown in Figure 5.

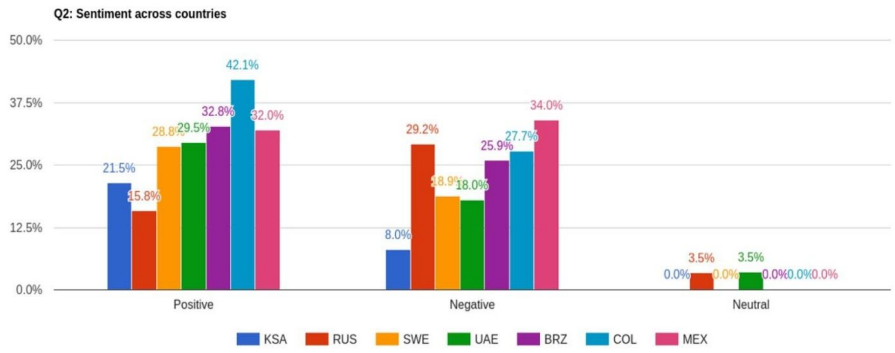

Figure 5: Attitude towards ”feminism” in seven cultural areas.

[9] http://www.si.se

[10] The full report (in Swedish) is available from the Swedish Institute: https://si.se/wp-content/uploads/2016/12/Sverigebilden-Rapport-_om_synen_pa_-jamstalldhet.pdf; a slide deck summarising the main points of the study in English: https://old.gavagai.se/Gender_Equality_Study.pdf

Explaining them by exploring the answers we found that feminism was associated with negative gender behavioural patterns such as machismo or with reverse discrimination in Latin American countries and in Russia, whereas it was accepted as a label for progressive policies and viewed comparatively positively in Middle Eastern countries. Example quotes ranging from positive to negative attitude scores are given in (5). This clustering of the textual data made possible to find explanations for the differences in attitude across cultural areas shown in Figure 5.

(4) If a man or woman describes themselves as feminist, what would you think of that person? What kind of associations do you get? Is feminism positive or negative in your view? How would you describe feminism?

(5) a. "Feminism is a positive concept, as women previously were discriminated against (earlier the world was sexist) whereas now women also find positions in areas which earlier were considered to be only for men."

b. "Feminism is neutral until it has acquired a mass character."

c. "I have a neutral view on this topic as each individual has their own perspective, as for me feminism shouldn't exist in today's world and education system."

d. "I consider feminism to be negative that it is the opposite to machismo or am I wrong?"

"What do you most wish for the coming year?"

In order to better understand their customers' thoughts and wishes for the coming year, AMF – a limited liability life insurance company owned jointly by the Swedish Trade Union Confederation (LO) and the Confederation of Swedish Enterprise (Svenskt Näringsliv) – sent out a survey to more than 100 000 senior citizens with 14 793 responses.[11] The survey included the open-ended question:

[11] The study is presented in greater detail by us http://gavagai.se/wp-content/uploads/2016/03/AMFPension-CustomerCase.pdf [13]The study is presented in greater detail here:
http://gavagai.se/blog/2017/04/24/what-makes-airline-passengers-happy/

(6) What do you most wish for the coming year?

Two thirds of the senior citizens responding to the open-ended question wished for better health for themselves, followed by concerns about their family, the global society and peace. The hopes were expressed in a manifold of formulations as might be expected from a broad sample of senior citizens from all walks of life. Clustering those into consistent topics would be a major challenge for any human operator, but with the terminology support we found handily that there were strong underlying topics in the content. Of the top ten topics expressed, three concern money and economy. In fact, make that four: the topic Utlandsresa (Travel abroad) also implies the spending of hard earned money.

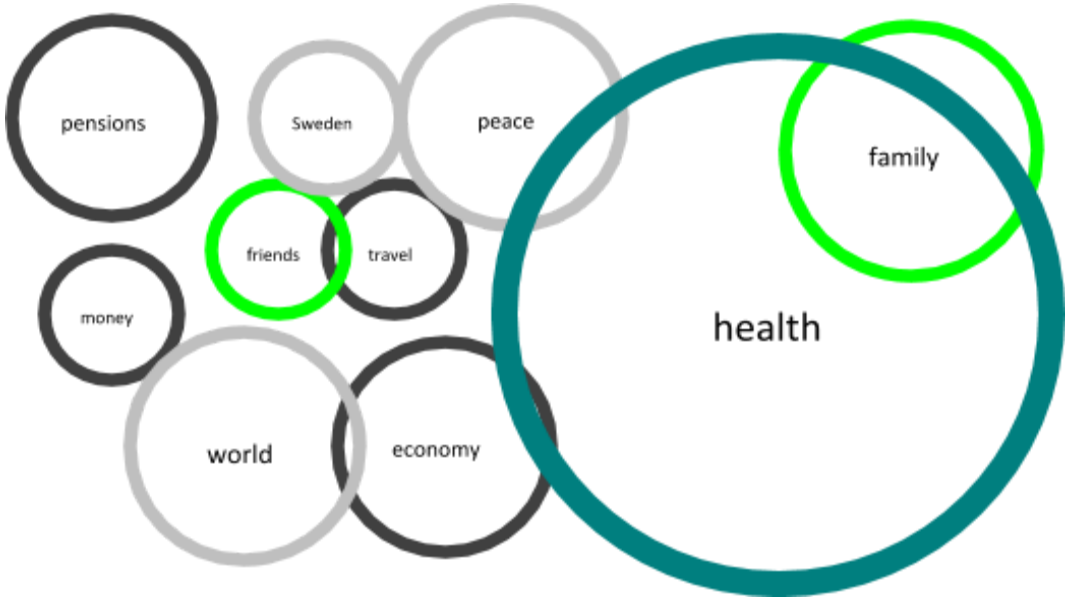

Figure 6: Hopes for the future among Swedish senior citizens.

”What makes airline passengers happy?”

We used the Gavagai Explorer to analyse online consumer reviews of airlines published on an online consumer review site.[13] We collected 20 000 free text reviews of 22 airlines, with no quantitative data attached to them

from the review websites. Attitude and topical topics are automatically identified and clustered. We measured how strongly opinionated reviewers are with regard to different aspects of their experience and we make these values comparable between different airlines. Some topics emerge from the text, with various degrees of frequency for different airlines: Food, Drink, Seat, Service, Value, Inflight Entertainment, and so on. Our main finding was that airline passengers seem to put up with almost anything, as long as they feel that they are being seen and looked after as individuals.

Passenger Satisfaction (descending %)	Satisfaction Score (%)	Areas for improvement (descending %)	Service Happy (%)	Service Unhappy (%)	Service Net Happy (%)
China Southern Airlines	0,62	Food (3,8), Service (3,1), Seat (2,6)	70	3,1	66,9
All Nippon Airways	0,53	Food (11,2), Seat (7,4), Service (4,9)	48,2	4,9	43,3
Lufthansa	0,43	Food (8,0), Seat (7,5), Entertainment	45,2	3	42,2
Qatar	0,422	Food (11,0), Seat (9,2), Service (4,5)	28	4,5	23,5
Singapore Airlines	0,395	Food (13,0), Seat (10), Entertainment	48,6	4,9	43,7
Norwegian					
Cathay Pacific Airways					
Thai Airways					
Qantas					
Emirates					
Turkish Airlines					
Virgin Atlantic					
Air France					
Southwest Airlines					
SAS					
Delta Airlines					
Etihad					
China Eastern Airlines	0,063	Service (24,3), Food (20,7), Seat (10,1)	18	24,3	-6,3
British Airways	-0,013	Service (19,2), Seat (18,8), Food (17,6)	27	19,2	7,8
Air China	-0,05	Service (23,4), Food (20,0), Seat (9,7)	19,6	23,4	-3,8
United Airlines	-0,152	Service (15,2), Food (10,0), Seat (9,3)	16	15,2	0,8
American Airlines	-0,275	Service (27,3), Food (8,4)	14,7	27,3	-12,6

Figure 7: Facets of attitudinal scores from passenger reviews of airlines, split on different airlines.

Data from our analysis are given in Figure 7. The satisfaction score (column 2) is generated by general-purpose sentiment analysis of the text.[12] The satisfaction score corresponds reasonably well with other known polls carried out in order to rate airlines for various top lists and awards.

[12] Karlgren et al. (2012)

The improvement topics (column 3) list the most pressing areas for improvement for each of the 22 carriers in the analysis. The happiest passengers (i.e. the passengers of the top-5 carriers in terms of or satisfaction score) complain to a small extent, and when they complain it is about meals (average 9.4% of reviews mention meals in a negative context), seats (average 7.3% mention seats in a negative context), service (average 4.1%) and sometimes about the on-board entertainment. The least happy passengers (i.e the passengers of the bottom five carriers in terms of our satisfaction score) complain mostly about service (average 21.9% of reviews mention on-board service in a negative context) and meals (average 15.3%).

The combination of sentiment analysis and topical clustering allows us to identify areas of improvement for the airlines, individually. This makes the results of review analyses much more actionable and shortens the path from attitude analysis to strategic business decisions.

Lessons learnt

The advantages of using automation for analysis are speed, consistency, and saving human effort for the most important tasks.

An automated system is quick: an analyst is able to analyze more answers in a given time period which means that the number of responses to a survey can be increased, gaining explanatory power. The marginal effort for a larger survey increases sublinearly.

An automated system is consistent: it will allow one analyst to process more data, and provides the same initial results regardless of who does the analysis, reducing the challenges of inter-rater reliability. If a coding scheme is retained for a survey performed repeatedly, e.g. every month, the analysis will not vary depending on who is coding the answers or the time between coding sessions for a certain coder.

Both of these factors reduce the cost and increase the reliability of the service.

You can try The Gavagai Explorer out yourself for free here: gavagai.io.

References

Douglass R Cutting, David R Karger, Jan O Pedersen, and John W Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 318–329. ACM, 1992.

Amaru Cuba Gyllensten and Magnus Sahlgren. Navigating the semantic horizon using relative neighborhood graphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2451–2460, 2015. URL http: //aclweb.org/anthology/D/D15/D15-1292.pdf.

Nick Jardine and Cornelis Joost van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5):217–240, 1971.

Jussi Karlgren, Magnus Sahlgren, Fredrik Olsson, Fredrik Espinoza, and Ola Hamfors. Usefulness of sentiment analysis. In *Advances in Information Retrieval*, pages 426–435. Springer Berlin Heidelberg, 2012.

Sofus A Macskassy, Arunava Banerjee, Brian D Davison, and Haym Hirsh. Human performance on clustering web pages: A preliminary study. In *KDD*, pages 264–268, 1998.

Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. A machine learning approach to building domain-specific search engines. In *IJCAI*, volume 99, pages 662–667. Citeseer, 1999.

Alicia O'Cathain and Kate J Thomas. " any other comments?" open questions on questionnaires–a bane or a bonus to research? *BMC medical research methodology*, 4(1):1, 2004.

Peter Pirolli, Patricia Schank, Marti Hearst, and Christine Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 213–220. ACM, 1996.

Dmitri G Roussinov and Hsinchun Chen. Document clustering for electronic meetings: an experimental comparison of two techniques. *Decision Support Systems*, 27(1):67–79, 1999.

Magnus Sahlgren, Amaru Cuba Gyllensten, Fredrik Espinoza, Ola Hamfors, Jussi Karlgren, Fredrik Olsson, Per Persson, Akshay Viswanathan, and Anders Holst.

The Gavagai Living Lexicon. In *Language Resources and Evaluation Conference*. ELRA, 2016.

Mark Sanderson and Bruce Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213. ACM, 1999.

Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846. Citeseer, 2000.