Paper 15

CRACKING THE CODE: WHAT CUSTOMERS SAY IN THEIR OWN WORDS

Tim Macer
University of Southampton

Mark Pearson
Egg Plc

Fabrizio Sebastiani
National Council of Research, Italy

the next fifty years
Introduction
This paper describes how Egg, a large UK internet bank, in partnership with Meaning Ltd, a research technology consultancy, commissioned the Institute for the Science and Technology of Information of the Italian National Council of Research (ISTI-CNR) to build VCS, a novel software solution for the automatic analysis of the many thousands of verbatim comments Egg collects through event and customer experience surveys conducted online. The sheer volume of responses received had made any systematic analysis of them impossible, which is a problem experienced not only by Egg but by many who conduct research online. These difficulties mean that researchers often restrict research designs to reduce or eliminate open questions from online surveys, despite the richness of the insights that can be obtained in this way.

Unlike other automated or computer-assisted software, the system developed for Egg is novel in that it applies machine learning to the problem of analysing and classifying verbatim texts from open-ended questions. With few precedents for the application of this method to verbatim textual data, which is characterised not only by great diversity of content but by frequent linguistic mistakes, a cautious experimental approach was taken in order to validate the software’s accuracy and reliability in performing the task.

The results of these experiments, presented in this paper, show that the software performs well against human coders that have subject matter expertise. Overall accuracy using a range of statistical measures was broadly similar. Furthermore, it was shown that accuracy could be improved by increasing the number of training examples, thereby demonstrating that the system is inherently trainable, and its performance can be monitored and improved with respect to each survey or question being analysed.

Practical experiences described here, since the software was adopted at Egg, demonstrate dramatic savings in cost and time, making comprehensive and highly systematic analysis of many thousands of verbatim responses cost effective.

It concludes by considering the implications of this technology for all engaged in research or evaluating customer feedback, and how its application could profoundly change the way research is undertaken, by effectively removing restrictions on placing open-ended questions in online surveys.

The Challenge
Open-ended questions remain a perennial feature of most quantitative surveys, yet extracting the meaning and value from them is usually the weakest link in the research process. The move to online research has exacerbated the problem, as participants tend to be more generous and expansive in their responses (Comley, 1996, Taylor, 2000; Smith, 2003). At the same time, online research raises expectation of near-instant reporting and actioning of all the results, which, until now, has been impossible to achieve where verbatim response data are involved. An important challenge of online research is thus analysing the large quantities of verbatim text returned by respondents in answer to open-ended questions, so as to allow prompt response both at an individual level and at an aggregate level.
CONTEXT: ABOUT EGG

Egg is one of the world’s leading online banks with over three million customers. It was launched in 1997 and offers a full range of financial services including credit cards, savings, insurances, mortgages and loans. It is currently a wholly owned subsidiary of Prudential. Customers need to be online in order to buy products and access services from Egg. Therefore Egg is able to deal with all of its customers online through its Website and by email. Egg does have a call centre used for servicing certain aspects of the products.

Central to the launch of Egg was its drive to be on the customers’ side and to redress some of the injustices towards consumers that it saw in the financial services industry. Egg declares its ‘Enduring Purpose’ to be:

“To revolutionize the customers’ experience of financial services driven through unleashing the power of people”

One of Egg’s strategic aims has been to help customers achieve the outcomes they wanted in their life using financial services products to do that. However consumers often made choices about such products without sufficient knowledge of whether the products purchased are the most suitable for them. Egg’s adopted strategy was:

“To transform how our customers feel about life by enabling them to progress what they want through making informed choices about their money”

Achieving a high level of emotional engagement with customers (e.g. their hopes, fears and aspirations) was a central part of this strategy. An important part of achieving this was not only through carrying out extensive customer experience and event-based surveys online, but crucially, in how Egg was able to respond to these both in aggregate, and at an individual level, when a customer was experiencing a difficulty, and had expressed this in a survey response.

The research team at Egg had developed an alerts system whereby they responded to customers who had expressed dissatisfaction through low scores in rating-scale questions; this was triggered from a closed survey question (Schwartz & Jennick, 2006). The next step was to interpret and categorize the emotions expressed in the thousands of verbatim comments that Egg received and ultimately generate alerts from the open-ended verbatim in the same way as had been implemented successfully for closed-survey questions.

MARKET CONTEXT

This section examines the context in which Egg and other similar organisations that interact with their customers through the Internet are operating in today’s business environment.

Technology

The World Wide Web today is a common part of people’s lives with over half of the population of the UK now regularly using it, according to Ipsos MORI MFS (2006). Consumers use it for entertainment, purchases, information and education, trading etc. and their consumption continues to increase with the advent of broadband and richer content (Web 2.0). And the future shows no sign of this relentless change slowing down: the convergence of hardware such as PDA’s, PCs and TV; the full impact of Web 2.0 and 3G and 4G mobile telephony. These developments are ultimately changing how consumers interact with companies.

As a result of new technology, conversations are increasingly taking place online – instant messaging, blogs and online communities are prevalent today. The research team at Egg, routinely observes from customer comments that they have a heightened expectation that interactions online share the same characteristics of these new conversational media. In particular, there are expectations that communications are no longer merely despatched from company to customer, but that that these communications are two-way, conversational — and virtually instantaneous.

Research

Research needs to offer consumers a dialogue. In the instance where the customer has just had an interaction with a company, not only does the company need to respond by seeking feedback promptly, but it must also be prepared to respond equally
promptly to resolve issues raised by the customer in the feedback provided. This was what motivated Egg’s research team to implement its alerts system linked directly to its continuous customer satisfaction survey and triggered principally by low scores. This has proved to delight customers, who are often surprised that the company actually reads returned surveys and is prepared to call and resolve the outstanding issue. However, responses constrained to a pre-determined range of answers in the closed questions of a typical online survey fall short of the ideal in a ‘dialogue’ where both parties to the conversation have the opportunity to express what they consider important.

The challenge for businesses

The challenge for both consumer-facing companies and research organisations is to meet the needs of both consumers and clients in this changing world. There is a plethora of tools to help do this such as word analysis tools, Web usage tools, and CRM technologies. What Egg was looking for was ways to have timely and targeted conversations through open-ended responses in its survey dialogues with customers.

In common with many other companies carrying out event-driven research, Egg pre-populates surveys with data extracted from its CRM systems. This includes items such as personal information, transactional information and, where appropriate, information about the classifications Egg uses to identify aggregate customer characteristics. The latter is extremely useful for reporting purposes as it is possible to correlate attitudinal information with transactional information, thereby comparing intention with behaviour.

Egg receives over 12,000 event-based and satisfaction-based survey responses each month. The alerts system described earlier was already providing a means to identify and respond directly to dissatisfied customers. But there remained a mountain of verbatim comments which the research team at Egg could only analyse very superficially – and this mountain was growing by around 20,000 more comments made by customers every month.

To respond to Egg’s strategic aims, the Egg research team had to find a way to make better use of this data. The scale of the task was too great for traditional methods, without vastly increasing the size of the team — which was not feasible. The same team was keen to find a solution which would automate the classification task, and thereby ensure that their work was focused on the more important task of analysing the results and helping to integrate change at a business and individual customer level.

Developments in the sphere of information technology in linguistics and automated text document processing mean that it is now feasible to deploy automated systems for survey research that will automatically classify verbatim responses to acceptable levels of accuracy, for analysis and reporting, segmentation or individual direct response. Yet when Egg, working in partnership with meaning ltd, surveyed the market for such solutions, it found that research technology providers were not engaging with these developments. Having invited a proposal to create a custom system from the Institute for the Science and Technology of Information of the Italian National Council of Research (ISTI-CNR) based upon advanced machine learning text document classification technology (Mitchell, 1996; Sebastiani, 2002), Egg decided to commission ISTI-CNR to develop a verbatim classification system (VCS) customized to its needs.

THE ART AND SCIENCE OF VERBATIM CLASSIFICATION

Coding or classification?

Though the technique we are discussing in this paper – the transformation of open-ended question responses into categorised data which may be analysed quantitatively – is most commonly called ‘coding’ in market research, this activity is more meaningfully described as classification or categorisation. However, as the term ‘coding’ is widely understood in the research community, we will use this term to describe the entire classification and categorisation activity performed by the VCS system.
The choice of software on offer

Prior to selecting the solution offered by ISTI-CNR, Egg and meaning ltd, working in partnership, researched the software marketplace extensively for suitable existing text classification tools that could be deployed to meet Egg’s need to classify open-ended texts in an automated fashion. In stark contrast to the situation with other market research survey activities such as online data collection or data analysis, where there is an abundance of competing products, we found this to be an area left virtually untouched by software developers.

While there were a number of solutions for processing verbatim responses, which are often built into a larger data collection suite of software programs, only a handful had engaged with the potential of machines now to read and understand textual documents. None of the built-in coding modules in any of the mainstream data collection packages used any automated text processing methods, and the majority had simply displaced paper methods which would have been familiar in any coding department in the 1950s with computer screen and keyboard.

Egg and meaning ltd approached those suppliers who appeared to be taking a more innovative approach to the problem of coding large volumes of data, in order to evaluate their solutions and select the eventual provider, and identified five types of existing solutions or potential solutions:

1. Manual coding by computer
2. Automated coding through word searching and text matching
3. Text mining or lexical analysis
4. Rule-based systems
5. Machine learning

During the early stages, Egg’s research team was open to adopting solutions from any of the last three classes of solution, provided that the solution could meet the design and implementation objectives that had been established. It was only after fully appreciating the differences and potential superiority of machine learning over the other methods that the decision was taken to focus on a this one method.

ANALYSIS OF THE ALTERNATIVE CODING METHODS

We will now consider each of the five different coding or textual categorisation methods outlined earlier, with reference to how each operates, the inherent advantages and disadvantages of each approach and the prevalence of software to support the method.

Method 1: Manual coding by computer

Through our own observation, this appears to be by far the most prevalent method in use today, and the method offered in most coding modules built into the standard and widely used data collection software products. Widely used products taking this approach include Confirmit, Voxco, Quancept, MR Interview, Snap and many others. We understand that some of these companies are now working on more sophisticated solutions; none were forthcoming at the time of our inquiries.

The method closely follows manual classification methods developed in the 1940s, often called post-coding or post-hoc coding, and is a two-stage process. In the first stage, a subsample of verbatim responses are drawn from the sample for each question. Topics and themes which appear to recur are identified and listed to form a ‘codeframe’ of answer categories. Typically, the final category will be ‘other’, to record those answers or parts of answers which do not resemble any of the identified answer categories. Subsequently, the actual coding process begins, and each answer is read in turn (including those used during the process of codeframe generation), and a ‘coder’ decides which categories from the list apply to that answer.
Advantages

This is a method without any notable advantages. The computerised form of this method may automate the production of the sample of answers, and eliminate the physical aspect of working with batches of questionnaires, as each answer from a CATI or Web survey will be delivered automatically to the coder for coding. However, these advances are relatively trivial.

Disadvantages

There are several major disadvantages with this method, to the extent that often online research design will omit open questions altogether due to the complexity of handling them, many of the issues remain unchanged since observations by Schuman & Presser (1979).

- It is very labour intensive, which also means it is expensive to administer.
- It is slow and adds a considerable delay to the process of preparing and releasing survey results.
- There is no economy of scale: the amount of effort increases in direct proportion to the scale of the task, as does the cost.
- Low levels of inter-coder agreement often exist but are not recognised (Montgomery & Crittenden 1979). While individual coders will often calibrate their coding work in the first instance, or coding supervisors may perform some spot check type quality control measures, the decision making process among coders inevitably differs widely. This leads to inconsistent, inaccurate results. According to Mullet (2003), coders may even exaggerate positive mentions or particular ‘acceptable’ positions. It is a problem which is exacerbated and is more difficult to reconcile as the scale of the task increases.
- The method is inflexible with respect to performing any kind of re-classification, for example, if a new category of answers emerges late in the coding process, or if any secondary analysis of verbatim responses is carried out which requires new categorisations to be introduced.

In the case of customer feedback, a critical flaw is that the slow speed of coding makes it impossible to target customers making a statement that indicates their relationship is in jeopardy for individual follow-up, and to do this in a sufficiently timely way that could lead to successful customer recovery. The issues of speed, cost and scalability were preventing Egg from making full use of the data it was collecting, and thereby inhibiting its ambitions to be in conversation with its customers, when some of those conversations were not being listened to.

No economic case could be made for recruiting sufficient manual coders to cope with the increasing volume of coding work, and the work offered would not be attractive or enjoyable work for those undertaking it, making it difficult to recruit.

Neither could placing the work with an offshore supplier be considered, as there were many important cultural and market-specific nuances in the texts which would be misclassified by anyone not immersed in the culture, and confidentiality and bank security issues also ruled it out.

Overcoming these disadvantages became the focus of Egg’s inquiries and ultimately the design objectives for the built software.

Method 2: Automated coding through word searching and text matching

This method has also become prevalent due to the very widespread adoption of Ascribe a proprietary Web-based coding solution for any survey data where typed-in verbatims exist, such as CATI, CAPI or Web surveys (Macer, 2002). Another earlier solution, Verbastat, used a broadly similar method (Macer, 2000).

In both these tools, the initial stages of codeframe generation and classification are merged. A coder works with the complete set of data, and then using a range of functions within the software, is able to group together similar responses by looking for
proxy words and phrases. Both offer support for handling common spelling and typing mistakes. Once an example answer has been found, searches initiated by the coder will seek out similar answers which can then be grouped together into a class.

At the end of the process, these groupings form the equivalent of a codeframe. However, in the background, all the individual answers have been tagged, and the coding scheme can then be applied to the data.

Some of the mainstream data collection tools also provide some rudimentary text searching capabilities within their coding tools. Bellview incorporates a ‘mass coding’ option, which will match and classify answers according to a search string. However, it cannot handle the complexity of full verbatim responses, and is only advantageous when handling questions with semi-open brand lists or similar items where answers tend to consist or one or two words.

Advantages

- Advantages of the method, when successfully implemented, include that it is:
  - Much faster, compared to method1.
  - Scalable: increasing the scale of the task, provided the questions are the same, only modestly increases the scale of the task for the coder.
  - A more engaging activity for coders
  - Addresses many consistency issues in two ways: often, one coder can handle all the data for a question or a whole survey; there is a complete audit trail of the coding decisions taken, and these can be undone or reworked at any time.
  - Offers some potential for re-classifying existing data.

Disadvantages

- Classification of continuous data remains a labour-intensive activity. Each subsequent wave of data must be re-processed. Although the classification scheme can be re-used, the majority of the work in decision making and categorising must be repeated for each new set of data.
- Direct intervention based on customers’ critical comments is difficult to make systematic.
- The method has only detected words or strings of words but has not assigned meaning to the words, which is critical when you are looking at re-contacting a customer with a problem.

Method 3: Text mining or lexical analysis.

Several tools are now available on the market which allow the user to interrogate and analyse text without the need to classify it first. They move away from simple text searching and pattern matching to methods that take into account the context of the words, and attempt to identify meaning rather than words. Solutions in this category include SPSS Text Analysis for Surveys1 and SPSS Lexiqueast, Provalis Wordstat and Sphinx (Macer, 1999, 2005).

Text mining relies on a combination of machine-based text interpretive methods usually based around dictionaries or lexicons which associate words or word phrases with synonyms or equivalent words and phrases which tend to share a common meaning. One commonly used method is Natural Language Processing, but there are others.

Most solutions tend to provide an existing set of dictionaries: those for cleaning up the text and recognising the word stems properly (a process called lemmatisation), and others for associating the words with others that share a similar meaning.
Advantages

- The principal advantage is that these systems remove the need to perform any categorisation at all.
- They make it much faster to deliver results from raw data.
- They are better equipped to handle exception reporting at a case level, where direct customer intervention is the outcome.

Disadvantages

- They increase the skill level of the person carrying out the task, as a skilled analyst is required rather than a trained coder.
- All the work must be re-done for subsequent reports.
- The data are difficult to interpret, and not amenable to reporting alongside the rest of the ‘closed’ survey data.
- They are heavily dependent on the creation of suitable lexicons or dictionaries. Meaning changes in different context (for example ‘low interest’ in the context of a survey on magazine readership and ‘low interest’ in a survey on credit cards would require entirely different treatments).
- This is an emerging area of development, and there is little consistency of approach, a leader in the field or significant traction in terms of customers using these products. This gives rise to concerns over future development and continuity of support.

Egg and meaning Ltd considered that the creation and management of these dictionaries would be a significant overhead in practice. It gives rise to a ‘black box’ approach whereby it becomes difficult to validate the system or understand how to correct systematic errors which may be taking place.

While we observe that in other situations this can be an extremely valuable tool to the researcher, in this situation it did not offer the potential for automation and reduction of manual effort being sought. As already noted, it would have required upskilling and recruitment of specialist resources to operate.

While this was not a disadvantage in Egg’s case, as all interviews are conducted in English, the system relies on having a set of dictionaries in each language. Processing interviews in another language will involve all the overhead of the initial setup again, which would be important for many researchers operating internationally.

Method 4: Rules-based classification

Another approach is to use rules-based text processing, where categorisation takes place according to logic determined by the user, an approach sometimes called an expert system. However, the system is able to identify meaning through the use of dictionaries and lemmatisation methods, as noted above.

One example of such a solution is iSquare’s i2 Smart Search.

Advantages

- This method has the major advantage that it can be automated to a very large extent, and therefore the scale of the task will diminish over time.
- Furthermore, textual answers are classified and their values can be recorded in the data, for subsequent analysis, as with conventionally coded date.
Disadvantages

- The number of potential rules to be written to classify the rather unruly set of responses that arrive from most internet surveys renders this solution unsuitable.
- Furthermore, maintenance becomes a programming task, which would require a different skill set from that available within Egg’s research team.

Method 5: Machine learning

The machine learning approach as proposed by the Italian National Council of Research (ISTI-CNR) takes a radically different approach from those outlined above. The methodology used will be discussed subsequently, so we will only outline the overall approach here.

With machine learning, the system learns by being provided with sample coded verbatim (a ‘training set’) which it then analyses in detail in order to establish common connections or points of congruence and cohesion within the examples provided (Giorgetti et al, 2003). Unlike other solutions which attempt to establish meaning by reference to either dictionaries or independent sets of rules, machine learning makes little attempt to understand what it is processing, but instead examines each new case provided in order to establish which answer or set of answers it shares most in common with, by reference to the words, phrases, and other linguistic patterns. The approach is broadly similar to that described by Raud and Fallig (1993) for applying neural network technology to the coding process. However, their approach differs from ISTI-CNR’s by placing an emphasis on unsupervised computer discovery of terms and phrases from the data in order to create coding schemes and codeframes automatically, rather than the human supervised training-by-example method proposed here. Indeed, a neural network is essentially a machine learning method.

Compared to the other solutions, a trainable machine learning system has the potential to overcome the majority of disadvantages of manual and other computer-assisted methods when it comes to text categorisation (Sebastiani 2002; Lewis 2003).

Advantages

- After initial training, the system would cope with any data provided now or in the future, with very little additional work.
- Training as a process was very similar to drawing up the codeframe and then performing manual coding using a conventional system. However, once sufficient training examples have been provided, the system could be turned on to continue any subsequent coding entirely automatically.
- It did not rely on dictionaries, so set-up required less effort.
- Retraining could take place at any stage, so that the system could cope with new classification categories or even changes in meaning over time.
- It also made it possible to re-categorise data for secondary analysis.
- Re-categorising historic and previously categorised data is possible and takes a matter of minutes to set up and run.
- It could fit in well with existing workflows, effectively substituting for the role of the manual coder, with the resultant classified data being stored in the same dataset as the rest of the survey, for routine analysis and reporting.
- It did not change the skill level of those required to operate the system.

Disadvantages

Two significant disadvantages existed. First, the system did not exist as a full-fledged software application. And second, although the method has successfully been applied to other types of data, such as syndicated newswire feeds, it had only been
verified for verbatim response data from surveys in an experimental setting (Giorgetti & Sebastiani 2003). Online surveys are bound to contain much more ‘noise’ and be much more diverse than data it has been proven to be effective with. The decision to implement a custom-built verbatim classification system using machine learning technology was not without its risks.

Machine Learning: Design and implementation objectives

For the automated coding system to be successful, and to integrate well with software, processes and resources at Egg, Egg and meaning ltd had identified the following key objectives that the new solution had to fulfil.

1. Once set up, tested and calibrated, the solution had to run virtually unaided on a continuous basis, with minimal intervention on a monthly basis.

2. The system should integrate with Confirmit, the Web survey data collection software used by Egg.

3. It should create standard coded data which could then be merged with the rest of the survey data for subsequent quantitative analysis and reporting.

4. The system should provide the means for researchers at Egg to verify that automated classification was being applied correctly, and if not, to be able to effect suitable corrections so that past and future data would be classified correctly.

5. The system should link to Egg’s alerts system, so that alert actions could be instigated directly from verbatim comments entered by customers.

Having examined the limited range of software products available on the market at the time our evaluation took place, in 2005, it became clear that the machine learning-based solution proposed by ISTI-CNR offered the greatest potential to meet Egg’s needs, and was the only solution which appeared to meet all five objectives.

In close interaction with both Egg and meaning ltd, ISTI-CNR developed VCS to meet the above design objectives, in consultation with those who would be using the software. The product now exists essentially as a set of interfaces between the proprietary coding technology developed and owned by ISTI-CNR, the survey data and the Confirmit data collection software, and the end users who manage and operate the software.

This highly modular design means that it can be re-deployed against other data collection software, provided that, like Confirmit, it exposes its survey data and metadata through accessible application programming interfaces (APIs).

In the case of Confirmit, it was an advantage that the API is provided as a series of Web services which means that the two programs are able to communicate and exchange data securely and at high speeds across the Internet.

Indeed, the VCS software developed has successfully met all of the design objectives.

THE SOLUTION

VCS is a software system for the automatic generation of automatic verbatim classifiers. A verbatim classifier is a software module that automatically determines which verbatim responses returned to a given question should be coded under a given code. A verbatim classifier is thus a ‘binary’ classifier, since for each verbatim the classifier takes a binary (yes/no) decision for the code to which the classifier is associated. For a given codeframe, VCS thus generates as many classifiers as there are codes in the codeframe (with the exception of the special code ‘other’, for which no classifier is generated).

How the VCS works

VCS generates a binary classifier for a given code after being exposed to sample (manually) coded verbatims that have been returned to the question associated to the codeframe. VCS uses these ‘training’ verbatims to ‘teach’ the classifier the
characteristics a verbatim should have in order to be coded under the code, pretty much as a parent uses photographs of tigers to teach a child what a tiger is. Actually, the parent will show the child both ‘positive’ examples of ‘tigerness’ (photographs of tigers, while saying to the child ‘these are tigers’) and ‘negative’ such examples (e.g., photographs of lions, panthers and so on, while saying to the child ‘these are not tigers’); it is by detecting the features that differentiate the former from the latter group of animals that the child learns to distinguish, from now on, tigers from non-tigers. Similarly, the classifier is trained to recognise positive from negative instances of the code by being exposed to known (i.e. manually coded) positive and negative verbatis that have been returned to the question to which the codeframe corresponds.

After all the classifiers for a given codeframe have been generated, they can be applied to new incoming verbatim responses. Each classifier will, independently from the others, attribute or not attribute its code to the verbatim; if no code has been attributed to a verbatim, the verbatim is coded as ‘Other’.

After verbatis have been coded by the automatic classifiers, a human coder may inspect the codes that have been associated to them. It is quite possible that the human coder may in some case ‘disagree’ with some of the classifiers, i.e. deem some of the code assignments (or non-assignments) wrong. This is quite natural: even expert human coders sometimes disagree among each other on whether a code should be assigned or not. In case of disagreement, the human coder may correct the code assignments performed by the classifiers, deleting erroneously assigned codes and adding codes that have been erroneously missed.

One interesting feature of VCS is that, in this case, the VCS can use these corrections to ‘show the classifiers their mistakes’: this actually means retraining the classifier to avoid these and similar mistakes in the future. In other words, the classifiers can learn from their own mistakes, once they are exposed to them. Actually, if the human coder instead agrees with the code assignments performed by the classifier and confirms this to VCS, this latter can use also this confirmation to reinforce the classifiers in their own judgments, thereby increasing the probability that future similar verbatis will also be coded correctly. This activity by the human coders is what we call ‘validation’ (of the automatically assigned codes). Indeed, validation may be seen as a way to provide the VCS with further training verbatis, which the VCS adds to the previously available ones in order to (re-)train the classifiers.

The ‘train and learn’ approach embodied by VCS is advantageous over the ‘rule-based’ classification approach, whereby an expert is required to manually write classification rules for each code or, globally, for each codeframe. With VCS, only providing a set of ‘training’ verbatis is needed; it is the system itself that takes upon itself the burden to generate the classifiers. This guarantees a high level of autonomy to the personnel of the customer satisfaction department of the company in administering their verbatim coding activities; in order to generate a classifier these personnel only need to feed VCS with ‘training’ coded verbatis, without recurring to the intervention of specialized personnel external to the company.

The same level of autonomy is guaranteed in case a new code needs to be added to an existing codeframe, or if a brand new survey needs to be set up; one simply has to provide the VCS with ‘training’ coded verbatis relative to the newly added code, or to the codeframes in the newly launched survey.

VCS is currently implemented as a ‘plugin’ to the Confirmit CRM platform. Confirmit does not itself provide an automatic verbatim coding functionality, and would normally require a human coder to perform the coding manually. VCS interacts with Confirmit by ‘simulating’ the actions of a human coder, fetching yet uncoded verbatis and depositing the coded verbatis back into Confirmit. This allows the user to profit from the rich reporting features built into Confirmit, in order to generate statistics and examine temporal trends resulting from the coded verbatis; from the alerting features built into Confirmit, in order to issue alerts to the company’s personnel if certain codes (e.g., the code ‘Customer may defect to competition’) are assigned.
CONTROLLED TESTS

In order to test the quality of the system ISTI-CNR ran several experiments on datasets of real verbatims collected by Egg in the context of its customer satisfaction surveys. We report here the results of these experiments. ISTI-CNR investigated two main dimensions: (i) the efficiency of VCS, defined by the time it takes the computer to build classifiers from training examples and then to apply these classifiers to coding yet uncoded verbatims; and (ii) accuracy, defined by the average rate of agreement between the coding decisions of the classifier and those an expert human coder would have taken for the same verbatims.

Experimental design

The system was tested on two datasets (hereafter named DS1 and DS2), generated by Egg coders under the supervision of meaning ltd, consisting of approximately 1000 manually coded verbatims each. DS1 and DS2 are relative to two different codeframes (CF1 and CF2) associated to two questions (Q1 and Q2) belonging to two different surveys. Two human coders (C1 and C2) took part in the coding of both DS1 and DS2. For both DS1 and DS2, the coding process was performed according to the following régime (incidentally, one routinely followed in most manual coding operations).

Firstly, coders C1 and C2 collaborated on coding the first 100 verbatims, so as to develop a common understanding on the meaning of each code. Subsequently, the two human coders coded a further 600 verbatims each, achieving an overlap of 300 responses classified by both, and two sets of 300 coded uniquely by each coder.

Of the resulting 1000 coded verbatims, 700 (comprising the initial 100 where the coders collaborated, plus each set of 300 coded uniquely by either coder C1 or C2) were combined to form the training set of responses, which was used to train the classifiers. The final set of 300 verbatims coded, independently by both C1 and C2, without any collaboration, formed the reference set. These were used for testing the coding accuracy of VCS.

The reference set were provided to the VCS classifiers without their codes attached, so that the classifiers would also code the data. This meant that the reference set had been coded independently three times: by coder C1, by coder C2 and by the VCS.

This provided a measure of triangulation, so that the performance of the VCS could be assessed against the naturally varying performance of two independent human coders. The variations in coding decisions could be computed along three axes.

The mathematical function used for computing this ‘degree of coincidence’ is:

\[
F_1 = \frac{2PR}{P+R}
\]

This represents the standard function for the evaluation of classifier accuracy in the field of automatic classification (Lewis, 1995), which corresponds to the harmonic mean of precision (P) and recall (R).

Precision measures the ability of the system to avoid errors of commission (codes attributed when they should not have), while recall measures the ability of the system to avoid errors of omission (codes that have not been attributed when they should have).

When F1 achieves a value of one, this signifies total agreement; whereas an F1 score of zero means total disagreement.

KEY FINDINGS

We will now consider the observed performance of the software at an overall level, with respect to accuracy and efficiency.

Accuracy

We have measured the agreement of VCS and C1, VCS and C2, and C1 and C2. The results are reported in table 1.
Table 1: Overall accuracy of the VCS compared to human coders

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Average number of training examples per code</th>
<th>Value of F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>C1 vs C2</td>
</tr>
<tr>
<td>DS1</td>
<td>50</td>
<td>.697</td>
</tr>
<tr>
<td>DS2</td>
<td>91</td>
<td>.764</td>
</tr>
</tbody>
</table>

This experiment (confirming the results of other experiments carried out by ISTI-CNR on different types of data) has allowed us to draw the following conclusions from the experiments:

(i) The accuracy of VCS correlates well with the agreement between different human coders.
   This means that if a code is one that C1 and C2 tend to disagree on (i.e., disagree whether it should or should not be assigned), then VCS will also tend to disagree with C1 and/or C2 (the table shows this tends to happen more frequently for the codes in DS1 than for the codes in DS2). This fact has at least two explanations. The first explanation is that, since some of the training verbatims had been coded by C1 and some by C2, this means that VCS has received somehow contradictory information on the meaning of the code, and this tends to lower the understanding of the classifier on the characteristics a verbatim should have in order to be assigned the code. The second explanation is that, if the disagreement between C1 and C2 is high, this means that the semantics of the code is controversial, and thus difficult to capture by an automatic tool.

(ii) The accuracy of VCS correlates well with the average number of training verbatims per code.
   This means, in essence, that the more coded verbatims human coders provide to VCS for training, the more accurate the classifiers are going to be (the table shows that the classifiers for the codes in DS2 are on average more effective that those for the codes in DS1). Coded verbatims may be provided (i) in the form of manually coded verbatims fed by the human coders before training begins, or (ii) in the form of automatically coded verbatims that are then validated by the human coders; there is conceptually no difference between these two modalities, since a verbatim that gets originally coded by the human coder and a verbatim that is automatically coded by the classifiers and is then validated by the human coder provide the same amount of information to the training process.

We have further tested how well VCS is able to approximate the true percentages of verbatims belonging to a given code. This is important, since customer satisfaction data are often important in the aggregate. For example, trends may be detected by comparing the percentages of customers that are coded under a given code in subsequent months; whether this percentage is growing may give important indications to the management on how the evolution of the product has impacted on the overall satisfaction of customers, and on how the product should then evolve from now on. Our test has been obtained by computing the pairwise standard deviation between our three coders (two human coders and the automatic coder) for each code in DS1 and DS2; lower values indicate better performance (a standard deviation of 0 will mean the two coders perfectly agree on the percentage of verbatims that should fall under the code). The results of the test are reported in table 2.

Table 2: Extent of agreement on the selection of codes by VCS vs human coders

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MAXIMUM STANDARD DEVIATION</th>
<th>AVERAGE STANDARD DEVIATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C1 vs C2</td>
<td>VCS vs C1</td>
</tr>
<tr>
<td>DS1</td>
<td>0.096</td>
<td>0.044</td>
</tr>
<tr>
<td>DS2</td>
<td>0.087</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>0.020</td>
<td>0.017</td>
</tr>
</tbody>
</table>
The results allow us to conclude that

(iii) The percentages of verbatims that the classifiers have coded under a given category correlate well with the percentages of verbatims that should indeed have been coded under this code.

Note in fact that VCS seems no worse than human coders at correctly identifying percentages. For instance, DS2 results indicate that, on average, VCS agrees with C1 more than C2 does!

The fact that detected percentages well correlate with actual percentages shows that aggregate data produced by VCS are reliable, and can be trusted by the management in their decision-making activity.

**Efficiency**

In terms of efficiency, VCS took roughly 4.1 seconds to generate a binary classifier on a standard 3.2GHz PC with 2 GB (average time across all codes in DS1 and DS2). Training times increase proportionally to number of training examples, number of codes per codeframe, and number of codeframes per survey; this means that, e.g., for a survey containing 5 open questions, each associated to a codeframe containing 20 codes on average, training from 1000 coded verbatims the classifiers for the entire survey would require less than 10 minutes.

On average, VCS took 1.3 seconds to code the 300 'test' verbatims. Coding times increase proportionally to number of verbatims to code, number of codes per codeframe, and number of codeframes per survey; this means that, e.g., for a survey containing 5 open questions, each associated to a codeframe containing 20 codes on average, coding the questionnaires of 100,000 respondents who have all answered all the open questions would take less than 40 minutes. ISTI-CNR is currently developing technology that will reduce these times by 90 per cent.

**Training performance**

We were also interested to learn the extent to which the performance of the VCS was affected by the quantity and quality of the data with which it was provided, with a view to understanding what an acceptable minimum number of training examples might be in order to create a robust classifier.

![Figure 1: The effect of increased training examples on accuracy](image)
ISTI-CNR run an experiment in order to assess the VCS ‘learning curve’; this consisted in testing the accuracy of the classifiers after training them with a portion only of the available 700 training examples, repeating the experiment for portions of increasing size. The results are reported in the accompanying graph; for example, DS1-test2 indicates the agreement of the classifiers (measured in terms of F1) with coder C2 as tested on dataset DS1, as a function of the number of training examples used (ranging from 100 to 700). The curves clearly confirm the above observation that the more coded verbatims human coders provide to VCS for training, the more accurate the classifiers are going to be. They also show that the improvement in accuracy is more marked in the initial phases of training: when only few training verbatims are available, providing more will markedly improve accuracy, while if many training verbatims are already available adding others will generate a smaller improvement.

PRACTICAL BENEFITS FOUND IN OPERATION

Cost and time savings

The new software, in its first few months of operation at Egg, has delivered approximately £82,500 of ‘coding value’ on its first project alone, for an expenditure of approximately four days’ work for a researcher in developing the classifiers through the training and validation modules.

This estimate is derived from its application to one survey containing approximately 225,000 verbatim responses. Coded manually, this would have required 3,750 hours of coding, at an assumed rate of 60 answers per minute. Using a junior coder paid at current market rates of around £11 per hour, plus 100% establishment costs, this equates to a cost of £366.67 per 1000 responses coded. For comparison, typical internal researcher cross-charges are unlikely to exceed £500 per day, or £2,000 in total.

It would take one coder working alone, two years and four months to complete the task, or more than 600 coders to complete the task in four days.

In practice, the majority of the work for those operating the system falls in the first month. Once a set of classifiers has been developed then subsequent waves of data can be classified with only a few minutes’ intervention required. Occasionally, a set of answers are checked in the validation module, which may result in an hour’s work for a researcher. If coded manually, each subsequent wave of 15,000 verbatim responses (from a monthly yield of around 10,000 completes for this project) would require a further 250 hours work: another seven weeks’ work for one coder. Currently, this is a little over an hour’s work for VCS.

At this time, no further systematic work has been carried out to verify accuracy, as performed during the test phase. However, subjective observation indicates that the data are as good as anything coded by human intervention. The member of the research team involved with the first survey used the word ‘uncanny’ to describe the system’s performance, commenting:

“It is quite spooky the way it classifies things. Sometimes it will code something you did not think of, but when you stop and look at it, you realise that it is often the VCS that got it right, and not you.”

Making better use of existing research data

In addition, within the first few months of operation of the VCS at Egg, the VCS is already facilitating a broader use of existing data. One example of this concerned a routine enquiry to the research team from a department who wished to understand what improvements it could make that would most benefit customers. In the past, this would have most likely required some new primary research. However, using the VCS, this research was handled effectively by performing secondary analysis on existing data – in this case, verbatim data from one of the main customer satisfaction surveys.

The request came from the team responsible for the look and feel of Egg’s customer Web site wished to make some improvements to the usability of the site. The brief was to identify ‘quick wins’ that could make use of a finite, short-term programming resource, and bring immediate benefit to customers. Using the VCS, the researcher was able to identify an existing classifier relating to Web site comments from one of the large, continuous surveys. It contained a general question about improvements, along the lines ‘what can Egg do better’. Using the VCS, it was possible to apply a verbatim classifier to categorise
those responses which referred to the website, and apply this category as a filter when exporting the verbatim comments. This had the effect of extracting around 350 specific comments about the Web site out of a pool of 27,000 verbatim comments.

Had this classifier not existed, it would only have been a few hours’ work to create one. However, as the classifier did exist, a report was provided to the client in 45 minutes from the time of the initial request. The report listed all the comments from customers relating to Web site issues from the most recent few months, sorted by date. It would be equally feasible to have organised this list by demographics or other sample characteristics, if relevant.

This basic report proved to contain many valuable insights for the Web design team, and made clear a number of areas where improvements could usefully be made. The verbatim classifier had been successful in identifying verbatim comments which related to web site usability in a sophisticated way that went far beyond what any simple text searching could achieve.

As a result, within the same day of the original request, the Web design team were able to review a manageable digest of relevant customer comments and take specific decisions about how best to allocate their development effort based upon what customers had reported. Previously, it is unlikely that this decision would have been informed by research at all, as the natural response of the research team would have been to collect more data, introducing costs and delays that would be unjustifiable for non-strategic projects such as this. Furthermore, the questions would most likely have been designed around a closed question presenting a list or menu of options derived internally. By contrast, the list of improvements that emerged was unfiltered by any internally held prejudices.

**THE FUTURE**

The continuous growth and capability of the Internet, greater customer demand for timely customer experiences and the way we all use research now will open the doors for a much wider use of the VCS technology.

The application of this technology effectively removes the logistical and economic barriers which have long existed in the administration of open-ended questions in quantitative research and which often results in them being designed out of surveys. By providing a ‘level playing field’ for open-ended questions against closed questions, we consider this technology has the power to transform the way quantitative research is conducted in the future, and lead to radically different and more daring use of open questions in surveys. It provides a new opportunity to hear the unfiltered voice of the customer in a systematic way, free from the influence of any pre-coded list of options.

Apart from having the system classify data automatically, this means we should also be able to look forward to segmentation models built on the comments and opinions given directly by respondents, as well as, or instead of demographic data and other proxy answers to closed questions.

Furthermore, by running the classifier in real time during an online survey, questionnaire routing and the logic of the survey could be interpreted dynamically and accurately from verbatim responses, rather than just from pre-defined categories, as at present.

As regards the software, currently, plans are being made for release 2.0 of VCS, which will contain novel features mainly addressed at further reducing human effort in the operation of the system; this will take the form of features for the semi-automated generation of codeframes, and for reducing the number of manually coded training examples needed for the generation of classifiers.

Going beyond the field of survey analysis, there is massive potential for the application of machine learning in the analysis of data generated from social networking, blogs and online communities. Here too, the amount of data is overwhelming, it is textual and highly unstructured. The software could directed be to the Web in order to categorise what is being said in these private-public spaces. It will allow researchers to identify then focus on salient online dialogues and, with relatively little effort, categorise analyse and capture the most important insights from these 21st century ‘man in the pub conversations’.
References


Footnote

1 SPSS have subsequently developed this product in version 2.0 to provide better categorisation capabilities. Our comments here relate to version 1.