Alan McMahon – Senior Consultant Analyst

View Profile Go to Blog

Lost in Translation?

8 tips so Multi-Lingual Text Analytics isn't too scary!

Were you ever faced with identifying insights from open ended text data? From call center notes, open ended questions on surveys, complaint letters, and/or online reviews? Have you tried this across multiple languages? Recently at Presidion, we had the opportunity to perform text analytics on multi-lingual telecommunications call centre data. We wanted to understand the most frequent issues our clients' customers where facing, and why they are calling the call centre. Between manually reading responses and using translators, the client had done very little with this rich information until now. This was mostly because they have customer comments in each of English, German, French and Italian and sometimes even more than one language in a single comment. If you've worked with multi-lingual text data before then you may know that turning this into something concise, consistent and understandable across all languages is quite a daunting prospect.

After initially separating the comments by the main language content, we were surprised how quickly we were able to make steady progress in identifying the major themes to report (called categories). The key to achieving a meaningful result, as with any text analytics project, was in understanding that the process is always subjective (whether manually performed or automated with text analytics software like the IBM SPSS Modeler Text Analytics we were using). Automated software brings with it many advantages (reducing person to person variation; a consistent analytical approach when repeated; summarising millions of responses relatively quickly) but it's not called unstructured data for no reason. We decide the structure we want to create, whether that be with hints from software or through business objectives and understanding. Here are our 8 tips to help you get the most out of multi-lingual text analytics projects.

1. Use a language expert! (but you won't need them as much as you think)

It goes without saying that having a knowledge of the language you are categorising gives you a distinct advantage, but you really would be surprised how far you can get without being fluent. For me it was enough to have two half days with a fluent language speaker during categorisation and one final day for review of the final results.

2. Separate the languages first and examine each separately

Create Text Analytics libraries as well as some general rules to identify the main language in each response. There may be no automatic way of knowing what language a conversation took place in and the assumption that a customer would communicate only in their 'preferred' language is, once tested, often not consistent with what you see in the data. The rules might be something as simple as "*if the text contains an umlaut then it is German*", or use the libraries to identify common words in a particular language and build this up from all the words/concepts found in the text. Be careful to look out for words common to more than one language though!

3. Expect to refine the results more than a few times

Text Analytics projects require multiple iterations of library creation and categorisation. Check the results each time and use the incorrect classifications to refine how each category is created. It is important to appreciate that the first solution is rarely the final solution, but there will also come a point where each extra hour you put into ensuring responses are correctly and completely categorised will give you less and less return. For example, reducing the incorrectly categorised responses in each category down to less than 5% or so might take you 5 hours, but the next 10 hours of effort might take you down to just 3% or 4%.

4. Set yourself short term deadlines

When you are performing the categorisation and refining the results, set yourself deadlines to prevent over doing it, like "*I'll work at this for the next 3 hours only, then do a half hour review*". This will stop you from chasing the one or two responses that are very difficult to correctly classify. Generally speaking, you want a high confidence in your categorisations being correct, but if only 5 or 10 out of 500 are wrong, don't waste too much time on that.

5. Start with only English (or the language you know best)

I found it helpful to do the entire categorisation process solely in English before beginning with the other languages. It allowed me to understand some of the main themes better and gave me a narrower focus. Different themes may present themselves in the other languages, but by and large the main themes should remain the main themes.

6. Focus on categories that are actionable (even if they are not as common)

When you are creating categories think about what can be done with this category – Does it imply a direct action that the client can take? For example, sometimes with call centre comments you will see some system generated messages such as "*Authentication failed*" or "*System Error – user not defined*". Often these are removed as they aren't direct opinions. However, if the system message indicates that the customer is experiencing difficulty it may point to a technical issue that needs to be resolved.

7. Keep it simple and report all results in one language

Too many categories or too many themes merged together make the results a lot more difficult to understand. Consider that with 50+ categories an organisation is still only likely to concentrate on or be interested in the top 5 or 10 to start with. In future projects you can develop the categorisation as the process and outputs become more familiar. Collating all responses regardless of language into one categorisation in one language will help too.

8. Verify the results

All the results you create should be peer reviewed. Firstly, if it is in a foreign language, ask somebody who is fluent to sense check that the comments in each of the categories make sense. Similarly have a colleague review the categorisation in case you have missed something. A business user can also verify your results so that you can ensure that there are no misconceptions and that all the categories are usable and provide some value.

For this project we were able to combine results from customer comments in both German and English into 4 major category types. Each of these had around 5 subcategories that were of interest to the client and the results were very well received. The final categorisation process took around 1 week for both languages and can be repeated now in minutes.

In this instance, while the client had conducted several questionnaires to retrieve customer opinions prior to our analysis, these results helped them to quantify both the scope of issues in their customer base and the efforts of the call centre teams in dealing with the topics. In the coming months and years now, they have an automated framework to identify any new topics and monitor progress as a result of any initiatives they may implement. That, and they don't need any translations anymore!!

About Presidion

Presidion have operated for over 20 years and have been the pioneers in implementing cutting edge predictive analytics solutions with top UK and Irish organisations. We specialise in helping organisations leverage their data to deliver tangible practical returns on investment, aligned with their strategies.

Presidion works with both government and commercial clients, currently partnering with hundreds of organisations enabling them to understand what has happened in the past, anticipate what may happen next to take appropriate and timely strategic decisions for their organisation.