NHS Staff Survey 2020 Free Text Analysis – Technical Document

About this Document

This document has been produced by Hertzian, text analytics specialists who worked with NHS England and NHS Improvement and the Survey Coordination Centre to process the free text data. The document aims to provide additional information and detail about the process used for analysing the data. It should be viewed as a document that can be shared with others involved in understanding the results of the analysis.

Platform Description

Hertzian's technology platform automatically analyses written text to discover topic and sentiment information within it. With Hertzian's platform, hundreds of thousands of comments can be automatically processed and its results can be used to discover key trends and insights within the data.

The Hertzian platform utilises Machine Learning to understand the context of the specific written text being analysed. This technology allows systems to be trained on data unique to industry (like healthcare) to ensure the context and nature of the discussion is understood and specific terminology does not cause the system to trip up.

Any piece of written text submitted through the Hertzian platform is analysed to identify the overall sentiment of the submission, any specific points of interest discussed and the sentiment corresponding to these interest points. The results of this analysis are then returned alongside the individual submission allowing for multiple forms of aggregation and analysis to take place.

Creation of Topics

As part of the setup and customisation stage of engaging with Hertzian, a unique set of topics were developed. These topics are used to present broad groupings of interest points and help provide an easier context for the overall findings from any broader analysis. The process of creating the topics involves the following three-part process:

- We first accumulate a corpus of written text data. In the case of the NHS National Staff Survey 2020, this was the entire response set for the two questions containing written text from all participating organisations. As part of this process, it's important to ensure that enough written text is compiled to cover the subject matter of the questions asked adequately.
- After accumulating our corpus of data, we train a Machine Learning model based on the corpus. This model learns the relationship and context of words. This produces what we call a "Semantic Map". The Semantic Map graphs words and their relationships, for example, a Semantic Map will understand the following:

  King – Man + Woman = Queen

  To assist with individual comment processing, each record of data is run through a process called "Phrase Extraction" which picks out all potentially relevant one, two or three-word phrases from the response. For each phrase, a positioning value is given by the Semantic Map.
- Once each record has been assigned a value, a process called "Clustering" is run on all of the data. The Clustering system takes these values, places them on a graph with each other and systematically begins to group them into groupings called "Clusters". These Clusters represent individual groups of terminology that are close together on the Semantic Map. For example, it will automatically group discussion referencing "appointments" & "consultations" due to the similar nature of the discussion taking place.

At this stage, the output of the automated system is handed over to the Hertzian team to conduct an in-depth manual review of the Clustering output. As part of this process, Clusters are placed into broader topics with several iterations produced to allow the client to share and relay key domain knowledge that aids with the placement of certain Clusters. This iteration process ensures the final topics contain as appropriate data as possible.

Processing

Once the above topic creation stage is complete, any new data that enters through the Hertzian platform can be assigned to the new topics. The following represents a typical journey that a single new record goes through as it enters through the Processing stage:

- The processing begins with the Phrase Extraction as outlined in the second phase of the topic creation process.
- A check is then run to determine if the identified phrase falls within any of the Clusters identified. If it does, it is assigned a label for that Cluster and the parent topic.
- If a topic is identified, the system isolates the context text surrounding the phrase and runs it through Sentiment Analysis which determines how positive or negative a piece of written text is.
- The outputted Sentiment (Tone) score and topic/subtopic found is then written to a file alongside the original written text content and any associated data such as Trust code, date etc…

If we take the following example response from the NHS National Staff Survey 2020 results, we can see an example of the various systems that run and the results returned:

> *"Working remotely has improved my happiness at work and my work-life balance. The flexible working makes me feel trusted, valued and I think I work harder because of it. A strong team bond has been formed during the pandemic, which I'm grateful for."*

The first stage of the Processing process identifies the following Phrases from the Phrase Extraction system:

- "strong team"
- "flexible working"
- "working remotely"
- "work-life balance"
- "pandemic"

Once checked against the Cluster results, the identified Phrases matched with mapped subtopics & linked topics:

- "strong team" -> **Team working/collaboration** subtopic -> **Working arrangements** topic
- "flexible working" -> **Flexible working** subtopic -> **Working arrangements** topic
- "work-life balance" -> **Work-life balance** subtopic -> **Health and wellbeing** topic

Finally, the record is then run through Sentiment Analysis to identify the Sentiment / Tone Score for each Phrase, resulting in the following scores for the corresponding segment of data:

- *"a strong team bond has been formed during the pandemic, which I'm grateful for."* -> Sentiment / Tone Score: **0.54**
- *"the flexible working makes me feel trusted, valued and I think I work harder because of it."* -> Sentiment / Tone Score: **0.41**
- *"working remotely has improved my happiness at work and my work-life balance."* -> Sentiment / Tone Score: **0.59**

Following processing any comments that remain unassigned to any topic undergo a further stage of processing. This secondary processing uses the identification of key terms (words and phrases) to increase the level of assignment and ensure maximum value from the process.

Outside of Scope

The following points describe what isn't possible with the Hertzian platform and some considerations to consider when reviewing the results:

- Upon individual review of each response, it may be that some comments are mis-tagged or placed into incorrect subtopics/topics, as with all automated systems, there is always a chance of this happening and efforts have been to ensure this does not happen frequently.
- Each comment run through the Hertzian platform has Sentiment/Tone attached that represents how positive/negative a statement may be. Due to the nature of how people leave feedback, there are some limitations to the Hertzian Sentiment Analysis platform. Difficult sentence structures such as ones containing double negatives or sarcasm may be incorrectly assigned a sentiment value.
- Some terminology may fall outside of the Clusters identified due to them being too similar to 'typical' language. Terms like "IT" and "room" may be too broad and have too much overlap with 'typical' terminology. However, efforts have been made to include these where possible in the subtopics and topics built.
- As the system works on a per response basis, should your record contain two or more fields of free-text, the system is unable to understand a link between the fields. In practice, this means that if a response to a question says "see previous", the system will treat this response individually and will not carry the sentiment/tone or any other data across.

Glossary

- **Clustering** – A Machine Learning process to group similar data without any pre-defined link.
- **Clusters** – The output of the Clustering process, contains a group of data that share similar traits.
- **Corpus** – A collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject.
- **Machine Learning** – Machine learning is the study of computer algorithms that improve automatically through experience. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.
- **Phrases** – The output of the Phrase Extraction, typically represents a specific interest area being discussed within the record.
- **Phrase Extraction** – The task of identifying the interest areas being discussed within a piece of written text data.
- **Sentiment Analysis** – The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic is positive, negative, or neutral.
- **Sentiment / Tone Score** – A score that represents how positive/negative a particular opinion is. Based on a scale of -1 (very negative) to 1 (very negative).
- **Semantic Map** – An output of the Hertzian model that identifies how similar certain terminology is with each other.
- **Subtopic** – Often contains one or a low volume of Clusters returned by the Clustering process.
- **Topic** – A manually compiled set of subtopics that make up a specific category of data.