Wave 2 2025: Methodology

Introduction

This wave followed the same quantitative analysis methodology as in Wave 1 of 2025. This combined Named Entity Recognition (NER), Surprising Phrase Detection (SPD), Link Analysis, Topic Modelling, and Semantic Mapping. We performed these analyses using Method52, a platform for social media monitoring and analysis co-developed by CASM Tech and technologists at the University of Sussex.

Where we have included quotes in order to elevate people's lived experiences, we have been careful to ensure that we preserve their privacy by bowdlerising the quotes. This method ensures that the meaning of the post has been preserved, but the wording or syntax has been changed, so that the author cannot be identified via text matching.

Quantitative methodology

Our computational text analysis proceeded through 3 stages: (1) Named Entity Recognition, (2) Surprising Phrase Detection (SPD), (3) Link analysis, (4) Topic Modelling, (5) Semantic Mapping and (6) data visualisation.

Named Entity Recognition (NER)

As in previous waves, we have used the technique Named Entity Recognition to extract names of individuals and organisations from our dataset. NER is a form of Natural Language Processing (NLP) wherein we use a pre-trained algorithm to identify particular entity types within a dataset. In our case, a model trained to identify people and organisations was used. For example, NER seeks to isolate proper nouns from within sentence structures, as these are likely to be of interest.

Surprising Phrase Detection (SPD)

SPD is a form of NLP which breaks down texts in a dataset into phrases and compares these with a reference dataset in order to identify which phrases are

more unusual[1]. It does so by analysing semantic features in each phrase and assigning a score for how likely the phrase was to appear in the reference data.

Link analysis

Link analysis is a process which extracts internet hyperlinks and website addresses from text. After extracting these links, we analysed which links appeared most frequently across the forums. Additionally, we analysed which website domains (such as www.gov.uk) appeared most within the links and how often they did so.

Topic Modelling and Semantic Mapping

Semantic Mapping (SeMa) is a part of the Method52 technology and methodology stack designed to aid analysts in performing quantitative and qualitative discourse analysis over a large number of documents. The SeMa process combines a human-in-the-loop strategy with large language model (LLM) based clustering tools, to organise each document into manageable, semantically related sub-collections. As a result, the analysts gain a better understanding of the source material whilst creating a reusable map of discovered themes for future work. SeMa can be broken down into two key phases: 1) a semi-automated clustering and topic analysis, and 2) an analyst driven thematic mapping, detailed below.

Document embedding and clustering

The embedding stage is the process of transforming each document into a numerical representation in a way that captures the document's semantics. Document embeddings can then be compared mathematically, such that documents with similar embeddings share similar meanings. Embeddings are computed using a pre-trained LLM.

For clustering the widely adopted approach of applying UMAP is used to simplify our numerical representations to a lower-dimensional space, suitable for the second step of applying HDBSCAN to identify clusters. We utilise the BERTopic package to encapsulate this process.

Thematic mapping

---

[1] Robertson (2019). Characterising semantically coherent classes of text through feature discovery. Doctoral thesis, University of Sussex. https://sussex.figshare.com/articles/thesis/Characterising_semantically_coherent_classes_of_text_through_feature_discovery/23470460; Robertson. 'Surprisingly Frequent Phrase Detection'. GitHub. https://github.com/andehr/sfpd?tab=readme-ov-file

Thematic mapping is the process of a human analyst inspecting samples of documents within each cluster identified in phase one and identifying whether that cluster is relevant to the research, and if so, applying a single theme and subtheme to that cluster. The themes and subthemes used for labelling are generated by the analyst, which will evolve as more of the clusters are understood and experienced. The process of applying labels to the clusters may require one or more iterations over the set of clusters as the analyst's understanding of the data grows. For example, master themes may be combined or split based on shared or disparately related subthemes. Conversely, subthemes may be removed, condensed or influence master themes as the nuances of the conversations in the dataset are better understood.

The aim of this process is to end with a coherent mapping of themes and subthemes to each relevant cluster, that provides a reduction of the clusters into a distinct set of semantic groups.

We then took random samples from the themes and subthemes identified by the thematic mapping, which we used for our qualitative analysis.

Visualisation

Finally, we visualised the results of our clustering and mapping using histograms of themes over time to present an overview of the thematic representation.

Qualitative methodology

Qualitative Content Analysis

In order to gain a more fine-grained understanding of our data, we selected the sub-themes based on our prior knowledge and the prominence of the subtheme in the dataset, as indicated by the number of posts labelled with the subtheme. We then generated a random sample of 100 posts from each sub theme. Additionally, we reviewed random samples of 100 posts which mentioned each of the following key terms identified through NER, SPD, and keyword analysis: Universal Credit, PIP, ESA, Liz Kendall, Rachel Reeves, and Keir Starmer. This was to ensure that we did not miss vital information and covered the most prominent conversations in the data.

Our analysts reviewed the samples using an iterative code sheet, which we adapted from Wave 1 of this year and expanded with additional codes where necessary. Our analysis involved tagging each post with any of the relevant codes. The result was an analysis of the significant areas of discussion within each of the sub themes, which showed where particular areas of discussion overlapped.