S4 - Topic modelling results

Explanation:

Coherence testing was conducted to determine the optimal number of topics to represent the dataset of food waste tweets. Topic coherence analyses the semantic similarity of the words within a topic and provides an overall coherence score, which determines the ideal number of topics in a dataset due to their meaning rather than statistical inference alone. Coherence testing was conducted on Python using the LDA method with Gensim implementation, based on the topic coherence pipeline detailed in the literature. The aim of coherence testing was to ensure a data-driven approach combined with expert knowledge to create meaningful topics. Coherence testing indicated that 13 topics (coherence score=0.476), 14 topics (coherence score=0.469), or seven topics (coherence score=0.460) were the most representative of the dataset. These topic models were created and assessed for semantic coherence by topic expert (ELJ) by looking at the topic distribution, keywords, and how similar the topics were to each other. Topic expert (ELJ) deemed 13 topics the most semantically meaningful for the dataset and discussed the decision with other topic experts (TAM, LB) to confirm 13 topics was most appropriate.

Coherence testing results:

The following table shows the coherence testing output for the topic model. 13, 14, and 7 topics were explored as options for the final topic modelling. Topic experts generated the three models and explored which was the most semantically meaningful based on the coherence score and their expert knowledge.

Table S2. Coherence testing results.

Topics	Alpha	Beta	Coherence
13	0.31	0.61	0.47628137
14	0.31	0.61	0.46911836
7	0.61	symmetric	0.46093161
7	0.91	0.31	0.46056594
12	0.31	0.61	0.45824344
8	0.91	0.91	0.45676227
7	0.61	0.61	0.45484672
8	0.61	0.31	0.45434198

7	0.91	symmetric	0.45422452
7	0.61	0.31	0.45154295

Topic model with 13 topics:

The following images show the topic model distribution and key words from each topic using the raw output from BigML. The topics had been named by topic experts, had the highest coherence score, and were deemed to accurately represent the dataset.

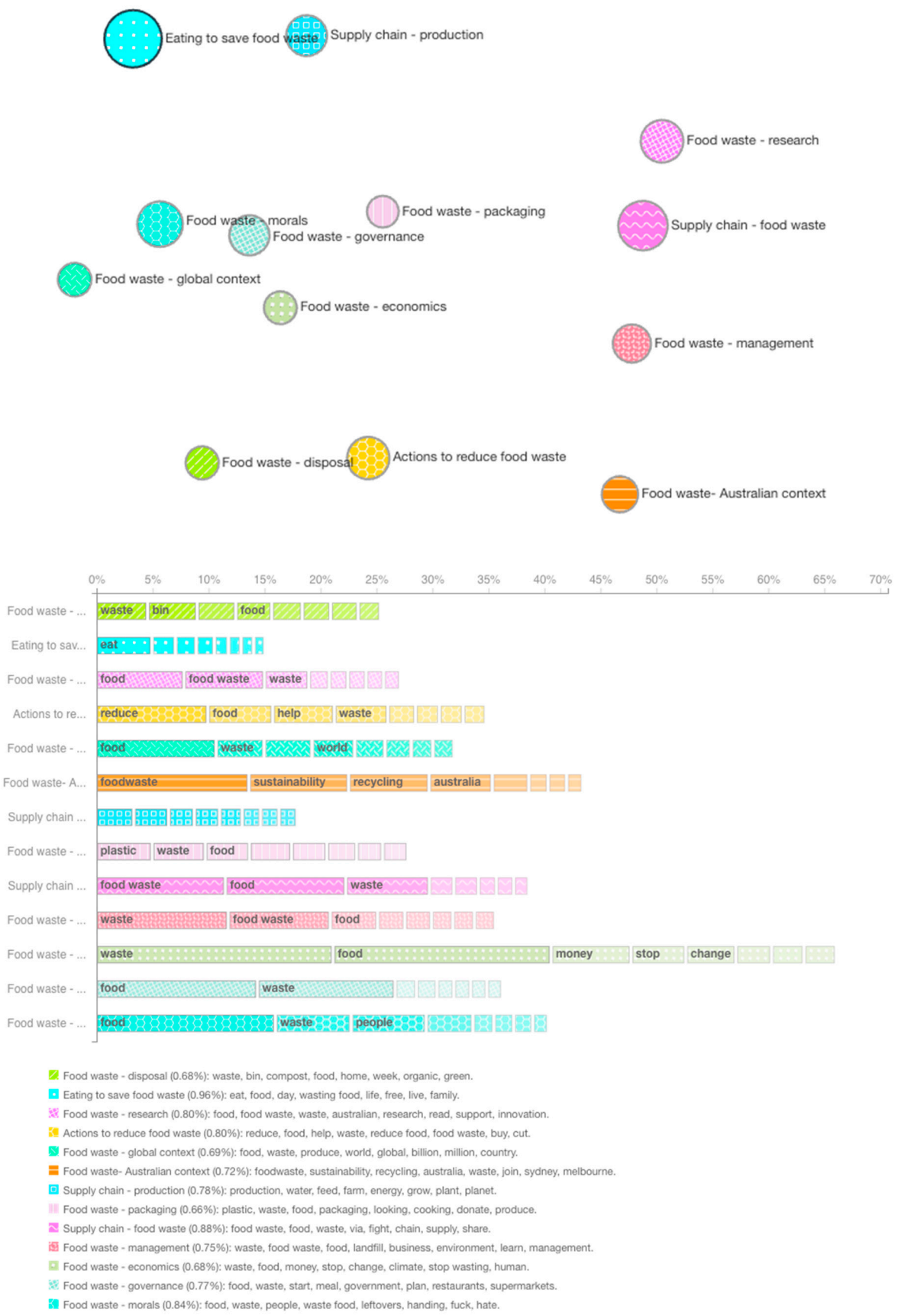


Figure S3. Topic model with 13 topics and keywords listed. The topics have been named by topic experts.

Topic model with 14 topics:
The following images show the topic model distribution and key words from each topic using the raw output from BigML. The topics had not been named by experts. 14 topics were not used as 13 topics had a higher coherence score and were deemed meaningful to the topic experts.

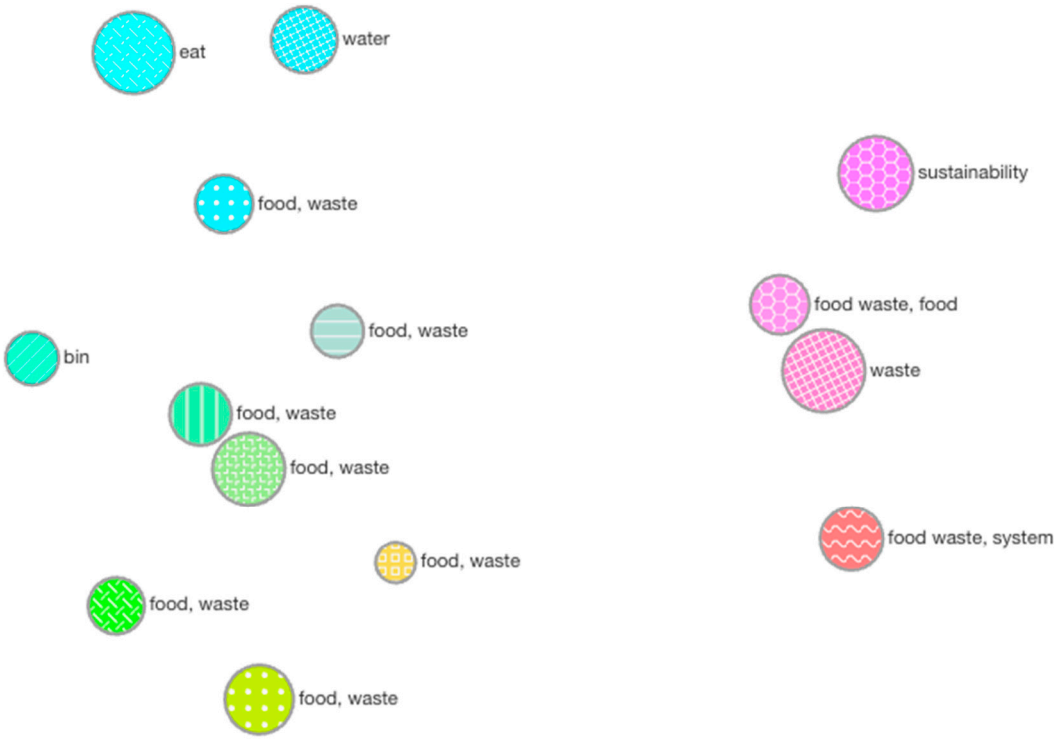




Figure S4. Topic model with 14 topics. Topics are given names by BigML.

Topic model with 7 topics:

The following images show the topic model distribution and key words from each topic using the raw output from BigML. The topics had not been named by experts. 7 topics were not used as 13 topics had a higher coherence and were semantically meaningful to the topic experts.

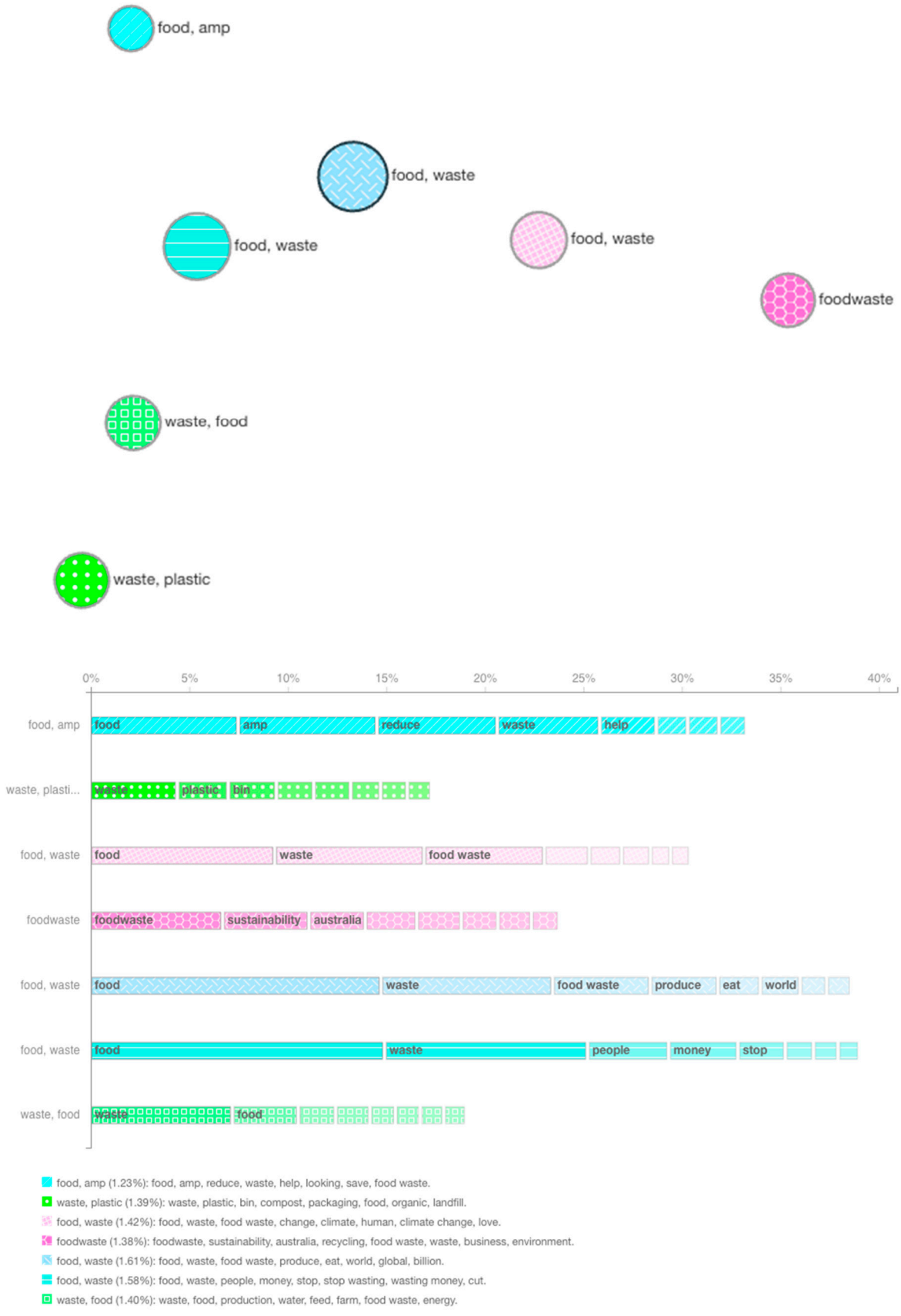


Figure S5. Topic model with 7 topics. Topics are given names by BigML.