Intellectual Property Office

Trade Marks Text Mining:

Topic Modelling of UK Trade Mark specifications.

Global Tech Mining Conference – Gareth Jones; 17th September 2024

Problem Statement (Slide 1 of 4)

In 2020 a proof of concept was commissioned to explore the UK IPO's trade mark data.

This identified an opportunity to gain additional insight into the trade mark landscape, which is a complementary indicator to patents and a relatively untapped resource for analysis.

Trade marks use the Nice Classification system, which describes the 45 broader classes of good and services.

The specification data however is completed by the each customer to describe the goods and services associated with their registered mark, so contains far greater insight.

Problem Statement (Slide 2 of 4)

Patents: International Classification (IPC) has ~600 subclasses.

Designs: International Classification (Locarno) has 241 subclasses

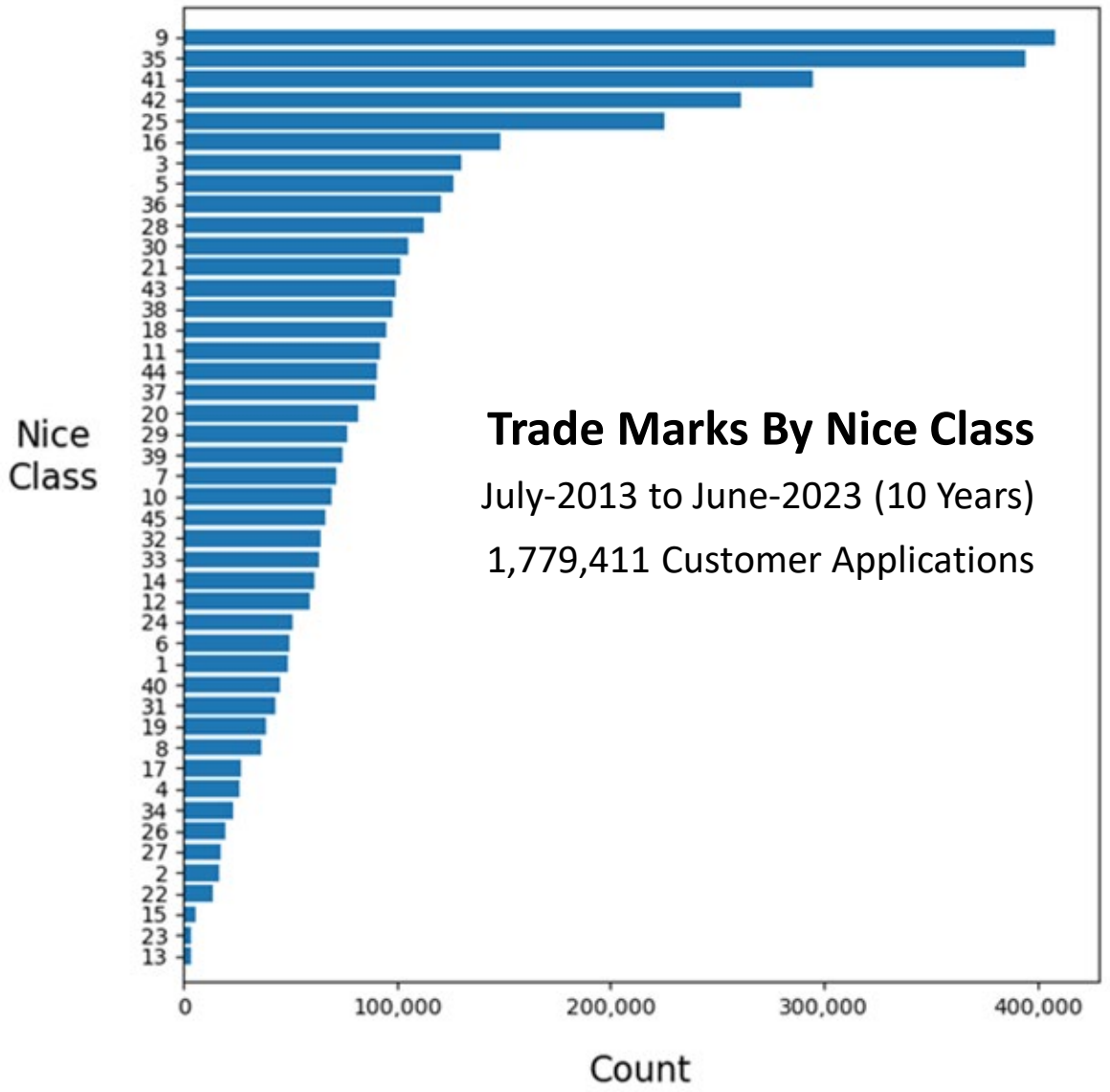
Trade Marks: International Classification (Nice) has 45 classes.

Intellectual Property Office

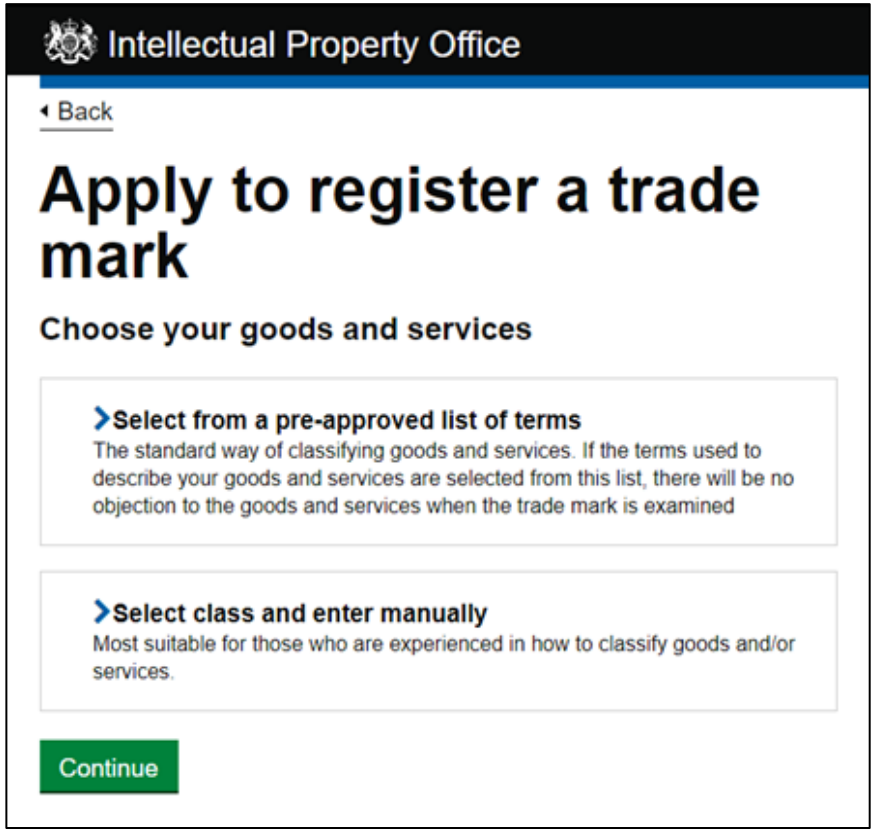
Problem Statement (Slide 3 of 4)

Class 9 (Technology) Header

“Scientific, research, navigation, surveying, photographic, cinematographic, audiovisual, optical, weighing, measuring, signalling, detecting, testing, inspecting, life-saving and teaching apparatus and instruments; apparatus and instruments for conducting, switching, transforming, accumulating, regulating or controlling the distribution or use of electricity; apparatus and instruments for recording, transmitting, reproducing or processing sound, images or data; recorded and downloadable media, computer software, blank digital or analogue recording and storage media; mechanisms for coin-operated apparatus; cash registers, calculating devices; computers and computer peripheral devices; diving suits, divers' masks, ear plugs for divers, nose clips for divers and swimmers, gloves for divers, breathing apparatus for underwater swimming; fire-extinguishing apparatus.”



Problem Statement (Slide 4 of 4)



Example Trade Mark Specifications

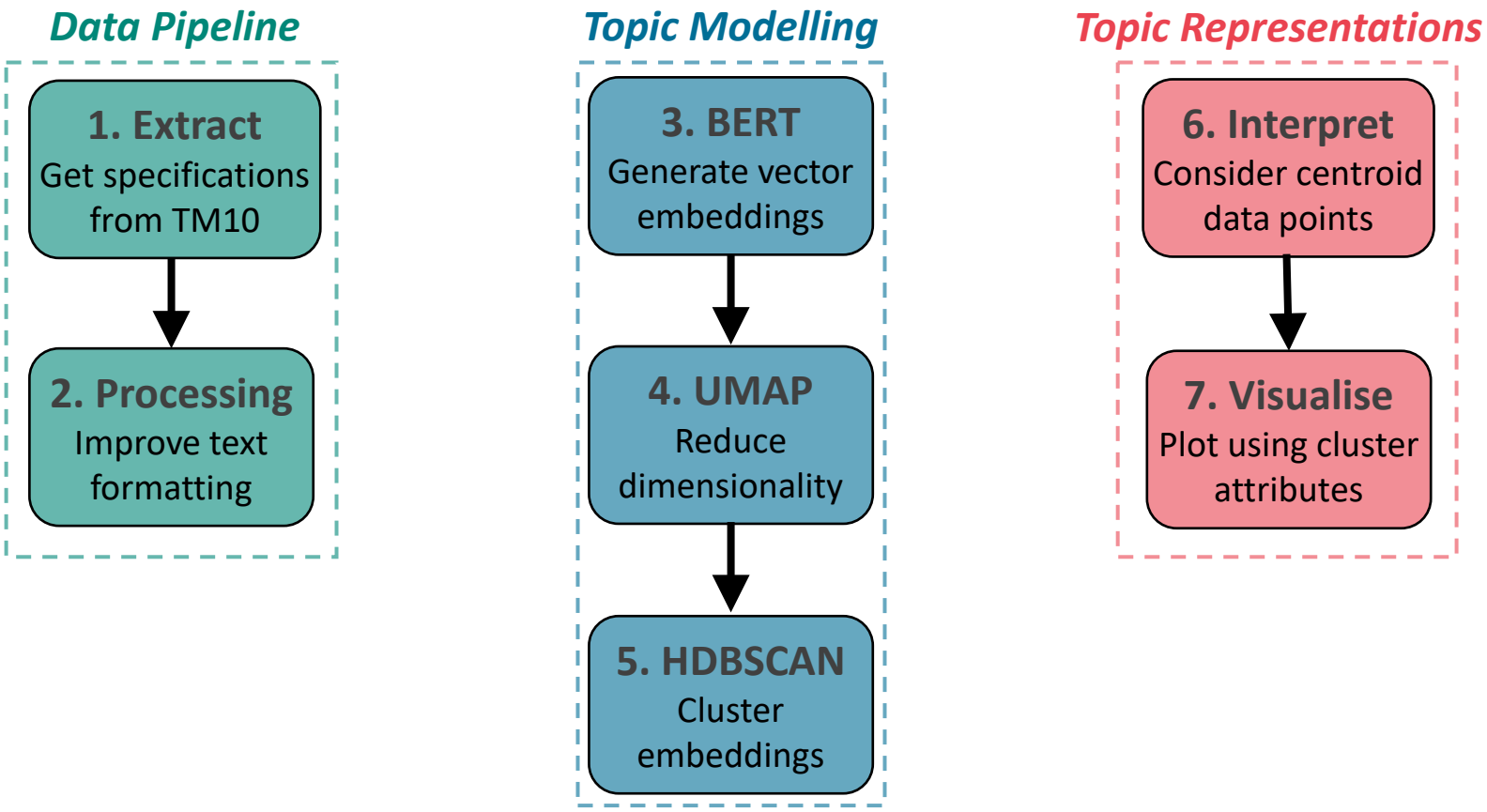
- Alcoholic beverages, cider and perry
- Distilled spirits, gin, vodka, rum
- Hats, jeans, shirts, tee-shirts
- Clothing, knitwear, sweaters
- Medical articles, Vehicles, Furniture

Data Limitations

- Mixed format
- Mixed Customer
- Varies substantially in length
- May contain one or many topics

Can we use NLP techniques on 'dirty' Specification data to gain insight into the Trade Marks landscape?

Methodology (Slide 1 of 2)



Methodology (Slide 2 of 2)

- Methodology taken from recent research (2022 paper)
  Article: Interactive Topic Modeling with BERTopic
  Paper: BERTopic: Neural topic modeling
  Github: MaartenGr/BERTopic

- Highest spec pre-trained sentence BERT model: ‘all-mpnet-base-v2’

- Use last 10 years of Trade Marks data, split by: Goods, Services, Class 9 (Technology)

- Preliminary work with One Hot Encoding and Word/Doc2Vec showed that BERT consistently produced the best clustering, using DBCV cluster evaluation metric.

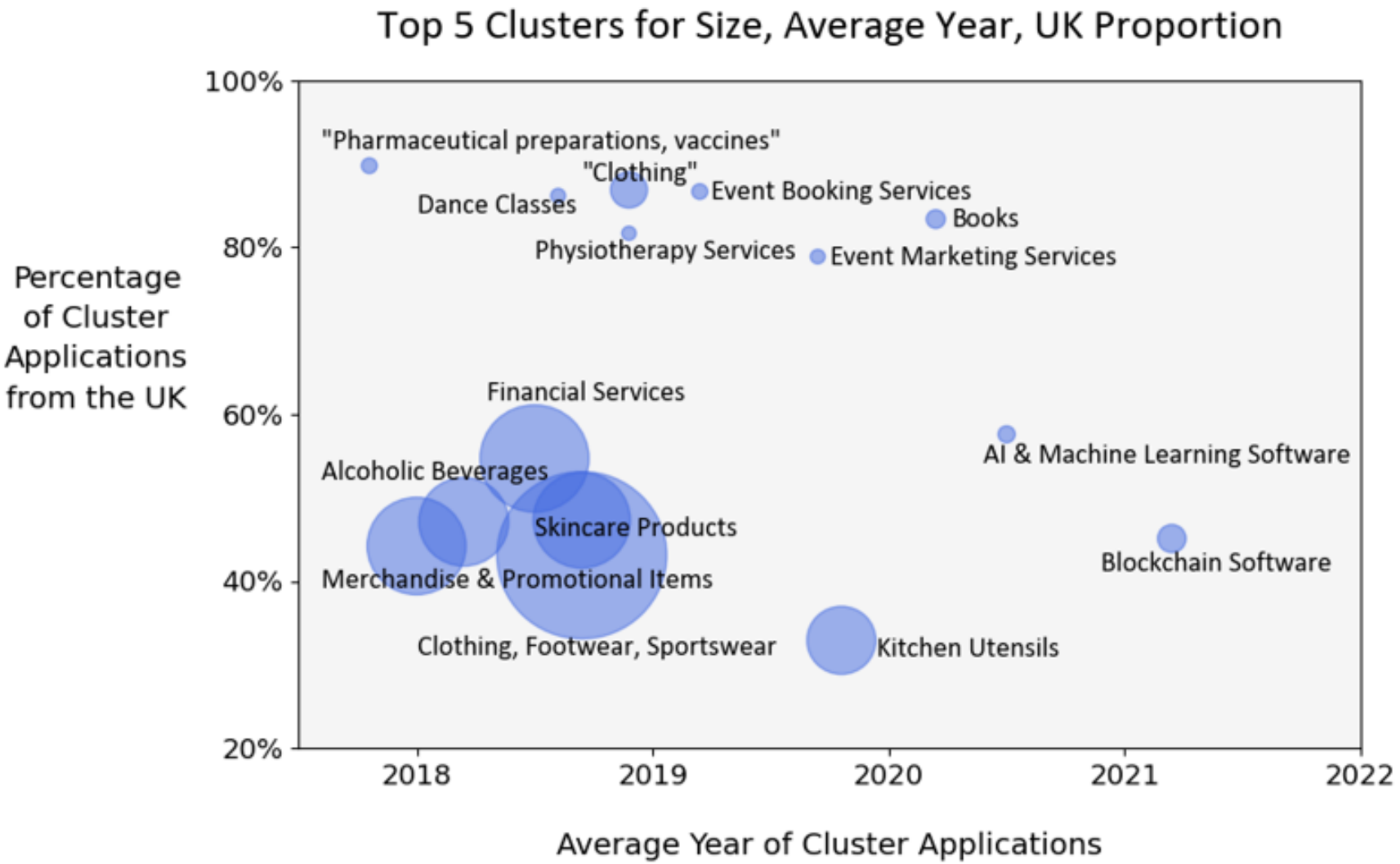
- Government guidance around the use and limitations of language models has been posted here.

Results (Slide 1 of 5)

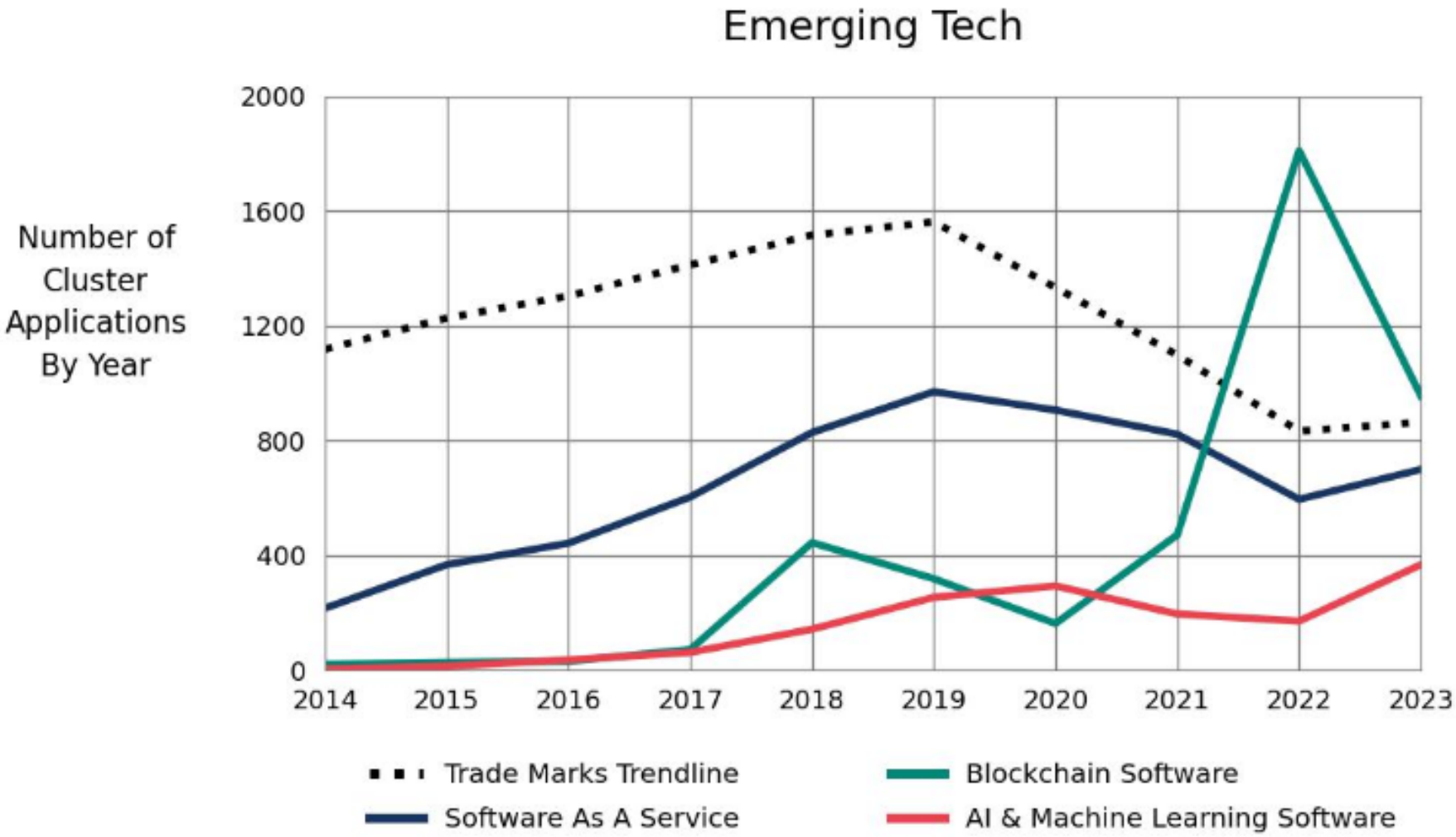
	# Clusters	Data Clustered	DBCV
Goods	113	67%	0.46
Services	92	61%	0.4
Technology	45	57%	0.37

	Data Not Clustered	OHE: 0	OHE: 1	OHE: 2-4	OHE: 5+
Goods	33%	6%	7%	9%	12%
Services	39%	0%	5%	14%	20%
Technology	43%	4%	7%	14%	19%

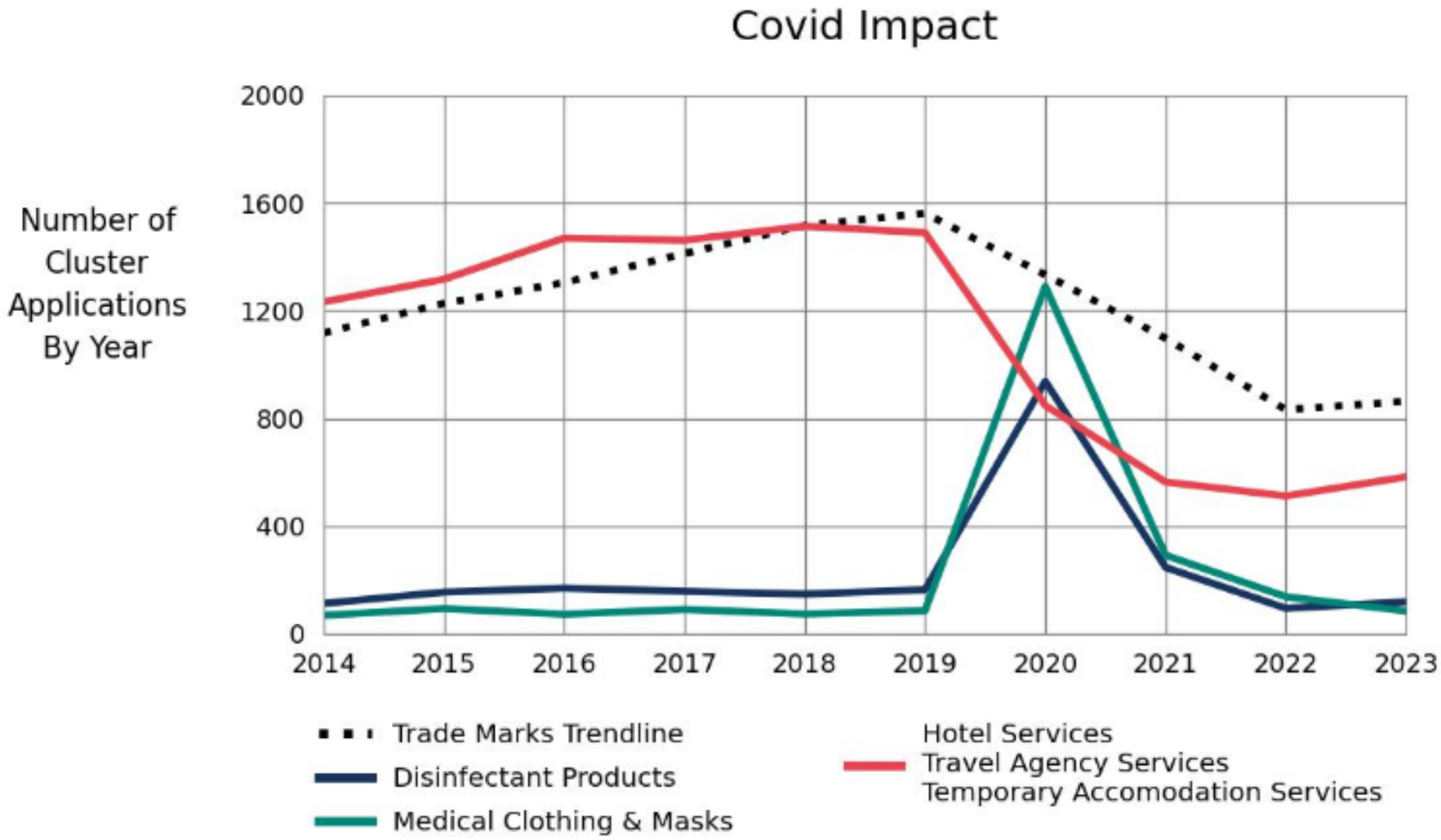
Results (Slide 2 of 5)



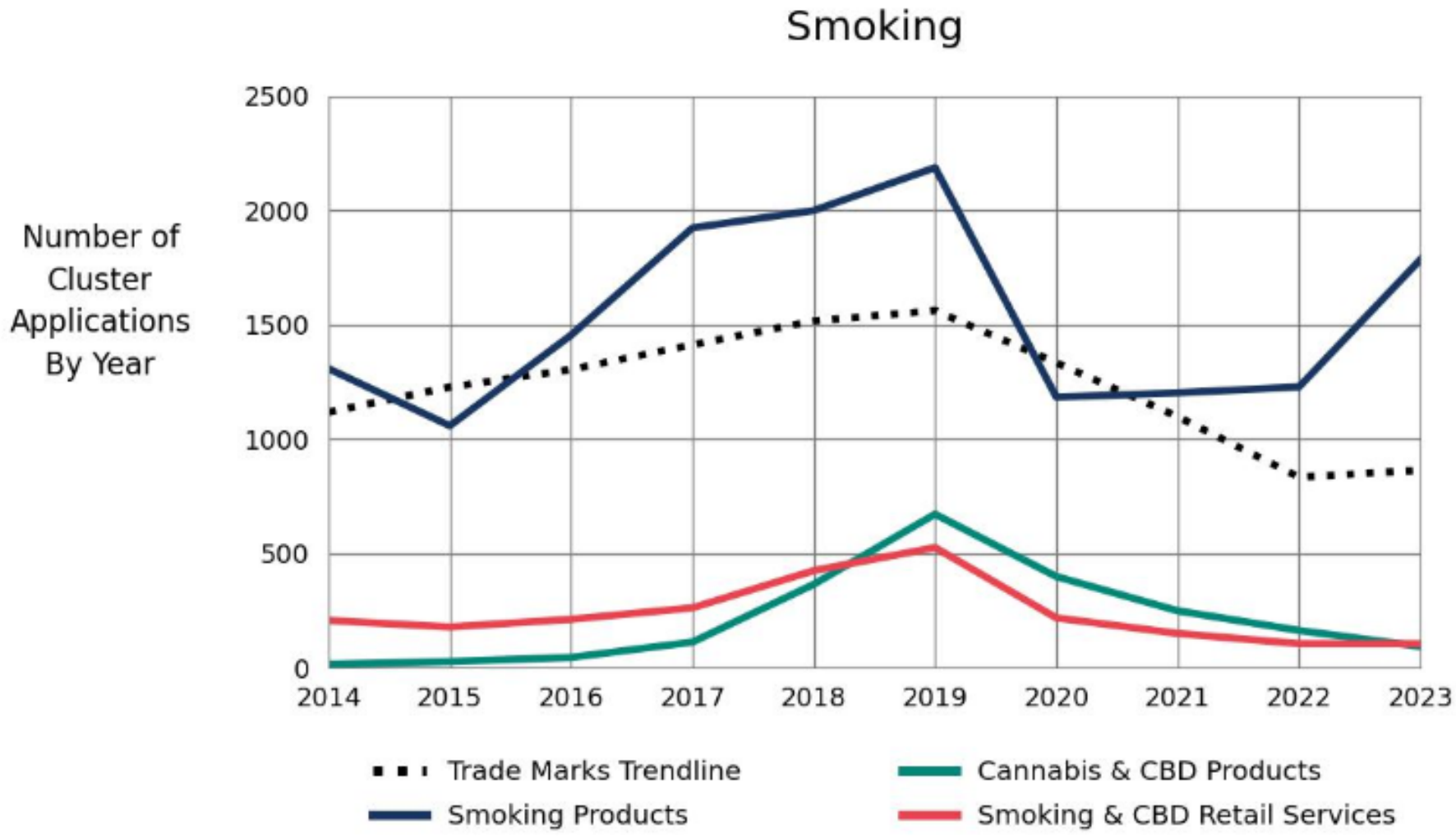
Results (Slide 3 of 5)



Results (Slide 4 of 5)



Results (Slide 5 of 5)



Conclusions

- The BERTopic methodology was successfully applied to UK trade mark specifications from the past 10 years to identify 250 topics

- This provides comprehensive understanding of the trade mark goods and services landscape

- Low level trends exist in the data for specific goods and services

- Further analysis is considered for technology areas such as blockchain, AI, Software as a Service, and gaming to understand how these fields have changed over time.

- Analysis could also be carried out to investigate CBD products in 2019 to understand the impact of government legislation

- Research and methodology to be published later this year

Thank you for listening