Using social media data to investigate the spatial and temporal heterogeneity in the perception of autonomous vehicles

3rd Symposium on Activity-Based Modelling

Technical University of Munich
10th December 2024

Eeshan Bhaduri
Charisma F Choudhury

ITS UNIVERSITY OF LEEDS

Existing AV studies

AV adoption & usage

Level of service

Attitudes & perception

1 Temporal variation

Values are
not constant

2 Geographical variation

Key research gaps

Stated preference approach

- Cross sectional
- Sample size
- Survey bias

Existing AV studies

AV adoption & usage

Level of service

Attitudes & perception

1 Temporal variation

Values are
not constant

2 Geographical variation

Key research gaps

Stated preference approach

Cross sectional

Sample size

Survey bias

How social media data fill the gaps?

- ✓ Passive user participation ensures **less survey bias** and **more natural opinion**
- ✓ Large scale data over long period aids in detecting **impacts of major events**
- ✓ (Potentially) understanding **sources of diversity** (social, institutional, political)

But it comes with its own set of challenges

- Potential **bias** in extracting sentiments from **short texts**
- **Non-representativeness** of the sample and lack of **socio-demographic** information

Research questions

How heterogeneous are the public sentiments towards AV?

a How have the **major AV events** impacted public sentiments over time?

- Identifying **peaks and troughs**
- Relation with major AV events and assessing **impact potential**

b To what extent do the public sentiments **differ across various countries**?

- Identifying **sentiment polarity** across **different countries**
- **Clustering countries** based on all three polarities

c What are the key themes of public discourse?

- Identifying the major AV related **concerns** and **enthusiasm** themes

Research questions

How heterogeneous are the public sentiments towards AV?

a How have the **major AV events** impacted public sentiments over time?

- Identifying **peaks and troughs**
- Relation with major AV events and assessing **impact potential**

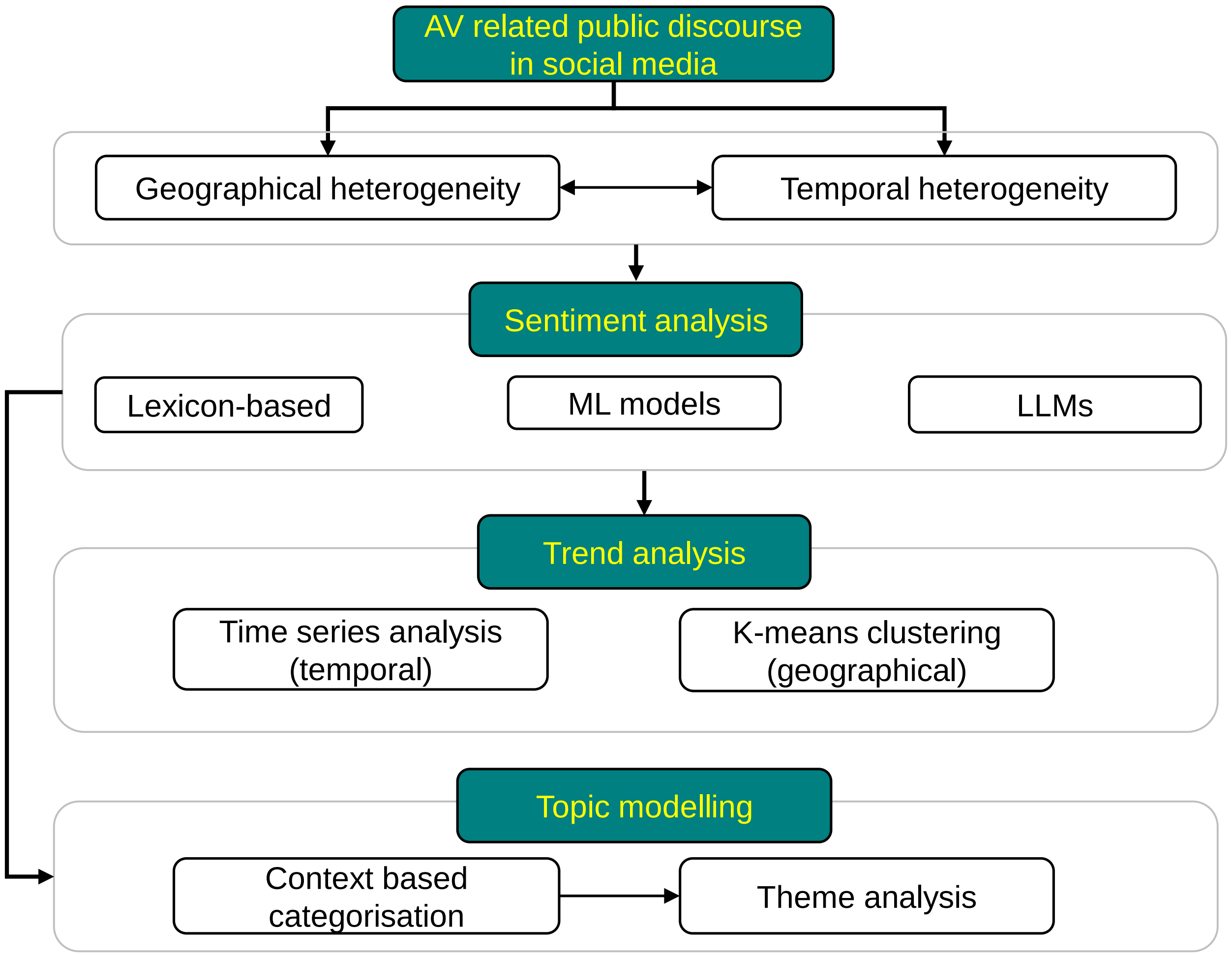
b To what extent do the public sentiments **differ across various countries**?

- Identifying **sentiment polarity** across **different countries**
- **Clustering countries** based on all three polarities

c What are the key themes of public discourse?

- Identifying the major AV related **concerns** and **enthusiasm** themes

Research framework



Step 0:
Selecting scope

Step 1:
Cleaning and filtering tweets

Step 2:
Understanding polarity of public sentiments

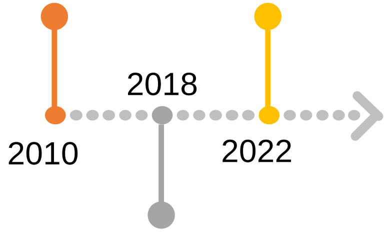
Step 3:
Uncovering temporal and geographical variation

Step 4:
Identifying major topics and themes

Descriptive analysis

Initial dataset: Used Twitter API to collect **600 thousand English language tweets**

1 Temporal variation



Time period: 12 years tweet dataset (2010-2021)

2 Geographical variation

Spatial variance: Filtered **91,429 geo-tagged tweets** from **11 countries** (at least with 1 million Twitter users)

English speaking: Australia (AU), Canada (CA), United Kingdom (GB), United States (US)
Lingua franca: India (IN)
High English proficiency: Germany (DE), Netherlands (NL), Sweden (SE).

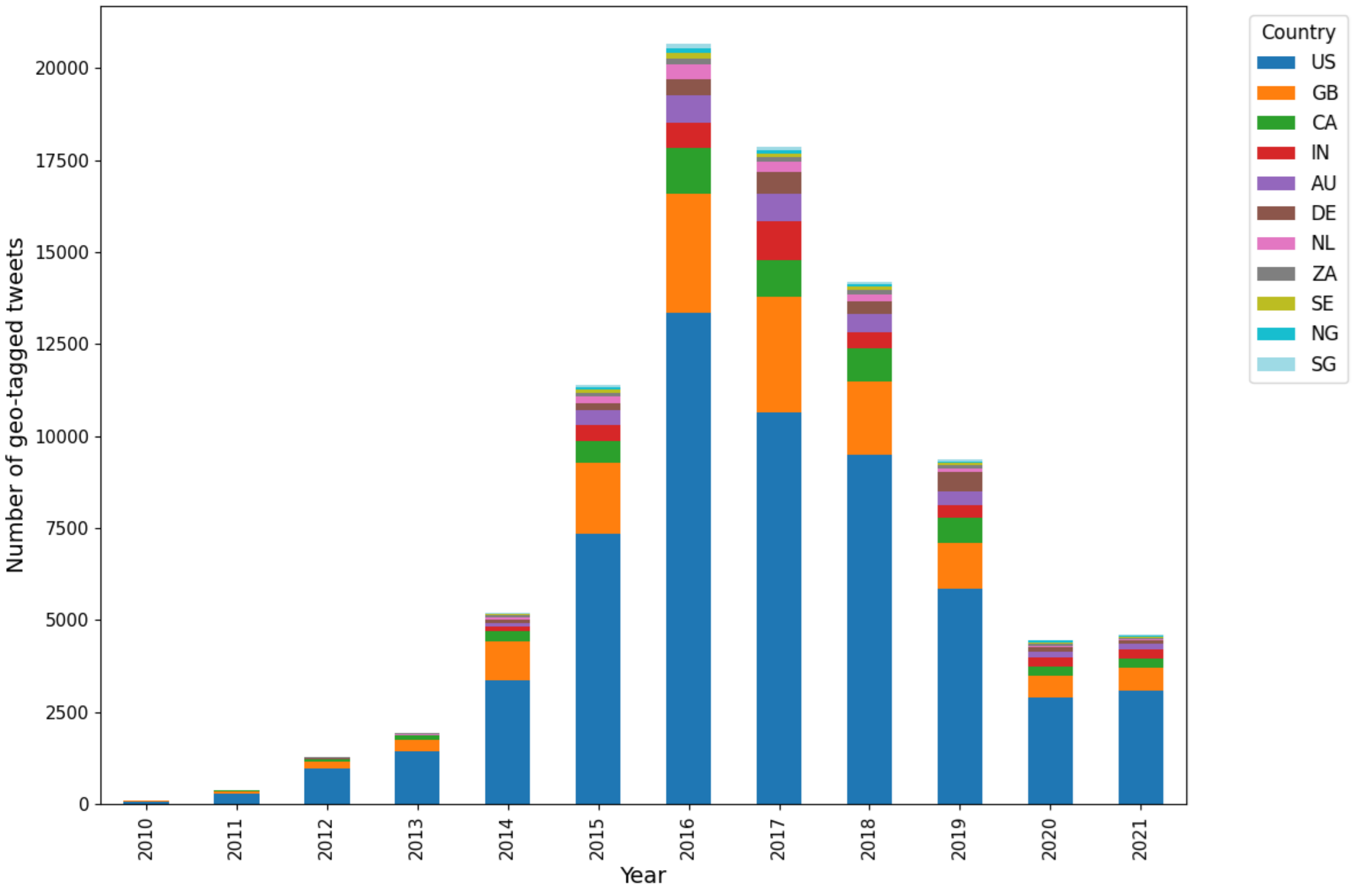


Fig: Number of geo-tagged tweets across study countries during 2010-2021

Sentiment analysis | Time series analysis | Clustering analysis

VADER: *Benchmark model*

Key advantages:

- **Pre-defined lexicons** (No training requirement)
- **Heuristics** based
- **Easier** and **Quicker** to implement
- Achieves nearly **50% prediction accuracy**

Table: Classification results of VADER

Sentiment Class	Precision	Recall	F1-Score	Support
Positive	0.5	0.48	0.49	214
Neutral	0.65	0.52	0.57	466
Negative	0.35	0.5	0.42	244
Accuracy			**0.50**	**924**
Macro Avg	0.5	0.5	0.49	924
Weighted Avg	0.53	0.5	0.51	924

VADER: Valence Aware Dictionary and Sentiment Reasoner

Fig: Word Cloud of key words in Twitter dataset

VADER: *Benchmark model*

Key advantages:

- **Pre-defined lexicons** (No training requirement)
- **Heuristics** based
- **Easier** and **Cheaper** to implement
- Achieves nearly **50% prediction accuracy**

Table: Classification results of VADER

Sentiment Class	Precision	Recall	F1-Score	Support
Positive	0.5	0.48	0.49	214
Neutral	0.65	0.52	0.57	466
Negative	0.35	0.5	0.42	244
Accuracy			**0.50**	**924**
Macro Avg	0.5	0.5	0.49	924
Weighted Avg	0.53	0.5	0.51	924

Why does not it suit our need?

- Inability to train for **context-specific words** and **sentiments**
  - Look out for words showing sarcasm and mixed emotions
- Limited ability to **counter class imbalance** and **no off-topic class**
  - **Off-topic** is minority class with 7.3% in annotated dataset

"Blameless self-driving car? Who is to blame, I wonder?"

"AVs are ssso totally safe!"

VADER: Valence Aware Dictionary and Sentiment Reasoner

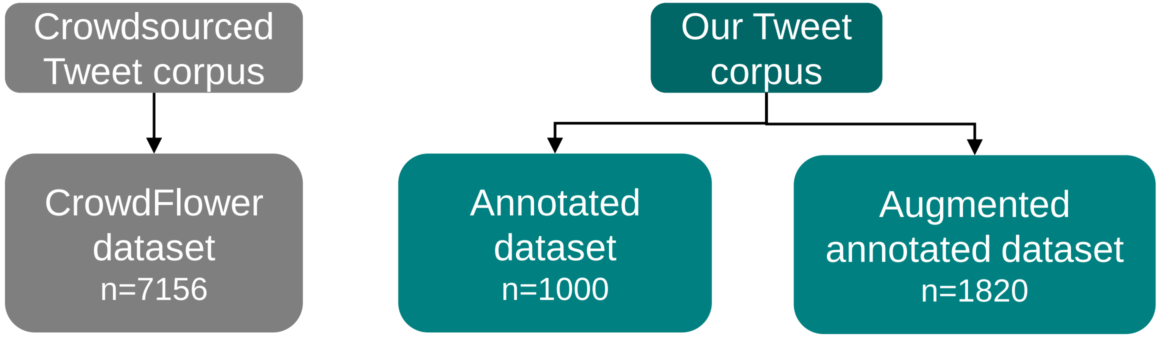
Fig: Word Cloud of key words in Twitter dataset

Machine learning models

Solution 1: Manual labelling

- **Works better** with a significant share of Tweets being **mixed emotion** and **off-topic**
- **3 annotators** achieving **Fleiss' Kappa** score **0.52**

Solution 2: Dataset augmentation



- **ML models** usually perform better with **larger training dataset**
- Augmentation (translation-based) provides **better data balance**

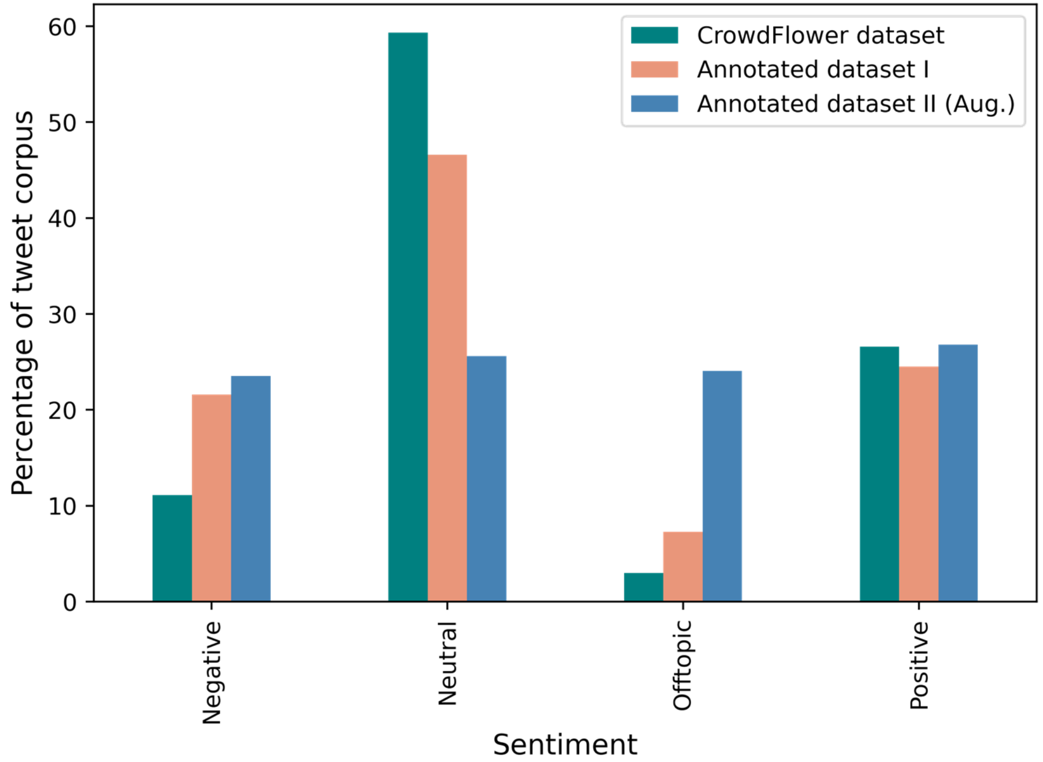


Fig: Frequency of sentiments in annotation datasets

Machine learning models

RF + SVM + NB

- Used **80%-20%** training-test split
- All 3 models perform best with **augmented annotation dataset**
    - **Data balance**
    - **Expert annotation**

RF: Random Forest
SVM: Support Vector Machine
NB: Naïve Bayes

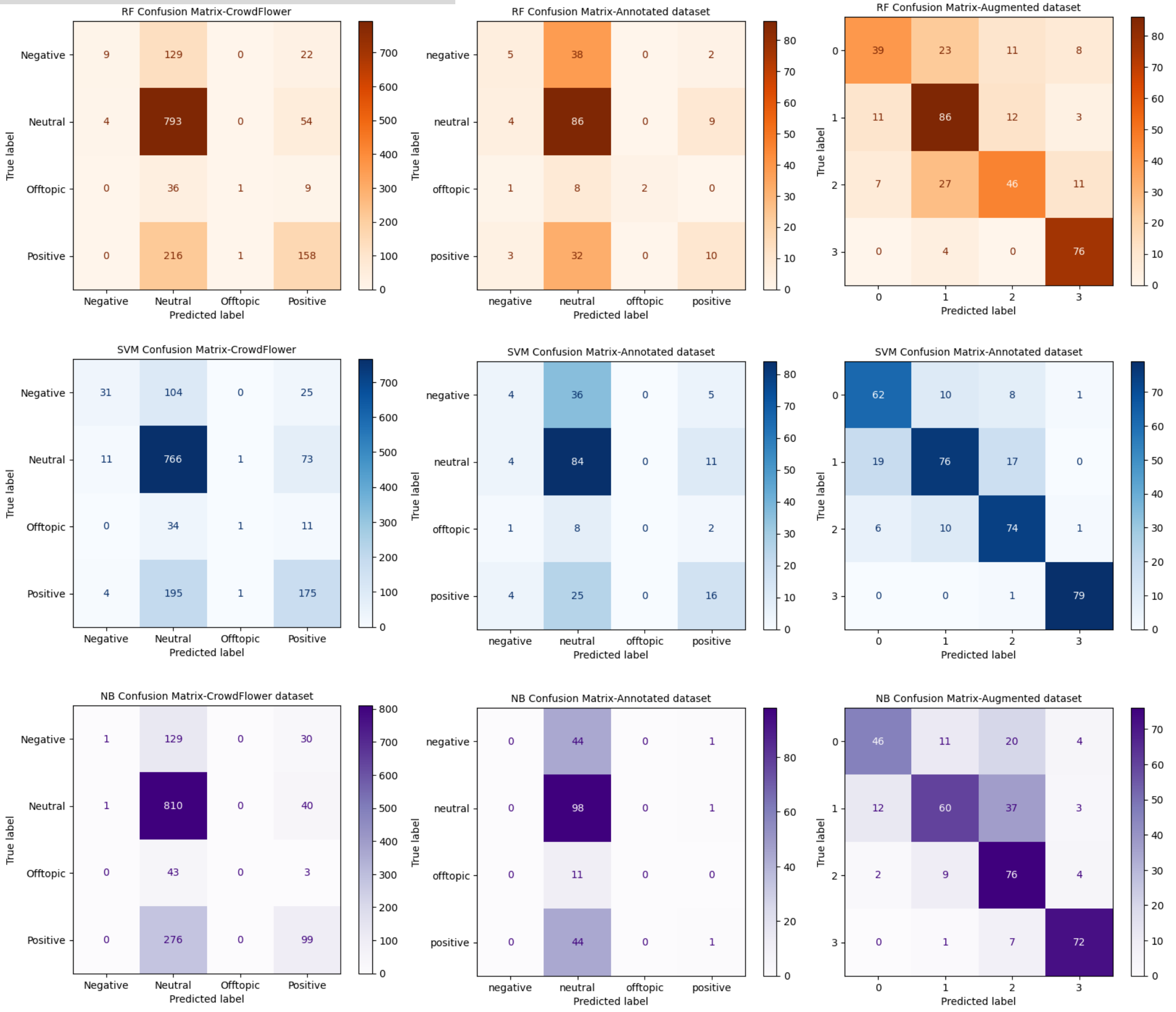


Fig: Confusion matrices of three ML models- RF, SVM and NB (from top)

Large language models: BERT

- Used **80%-20%** training-test split
- Performs better than ML models
  - **Fine-tuning** (annotations) on top of **pre-trained BERT model**
  - **Bi-directional nature** better captures dependencies (Masked language modelling and next sentence prediction)

BERT: Bidirectional Encoder Representations from Transformers (BERT)

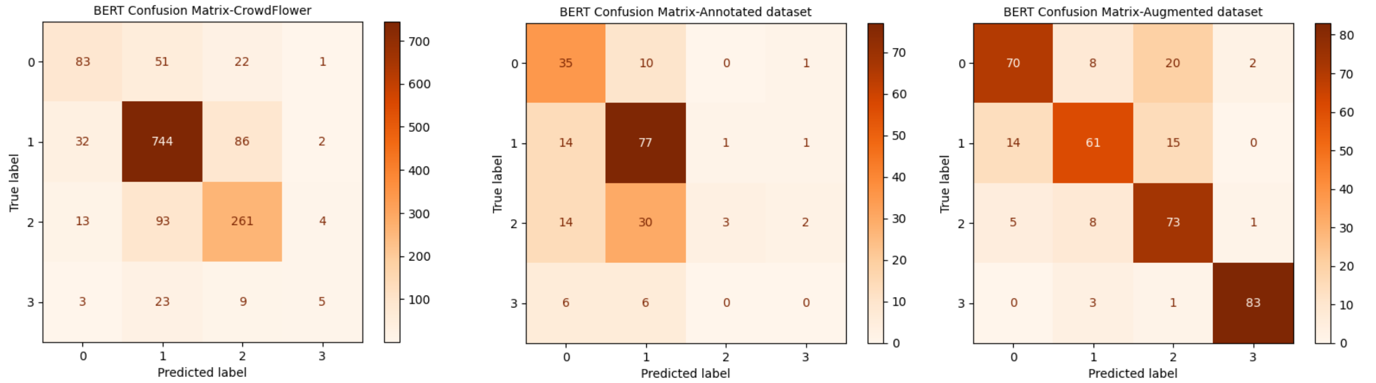


Fig: Confusion matrices BERT (for three annotation datasets)

Table: Classification results of BERT

	CrowdFlower data			Annotated dataset			Augmented annotated dataset		
Index	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Negative	0.634	0.529	0.576	0.507	0.761	0.609	0.787	0.700	0.741
Neutral	0.817	0.861	0.838	0.626	0.828	0.713	0.763	0.678	0.718
Positive	0.69	0.704	0.697	0.75	0.061	0.113	0.670	0.839	0.745
Off-topic	0.417	0.125	0.192	0	0	0	0.965	0.954	0.960
Accuracy			**0.763**			**0.575**			**0.788**
macro avg	0.639	0.555	**0.576**	0.471	0.413	0.359	0.796	0.793	**0.791**
weighted avg	0.753	0.763	**0.755**	0.592	0.575	0.499	0.795	0.788	**0.788**

- Captures the major events
  - Major crashes show **high impact of negative incidents**
- Indicates gradual decline in AV interest
  - **slow dying of interest** for AVs
  - **people getting more sceptical** after accidents
- **Highly dominated** by **USA events**

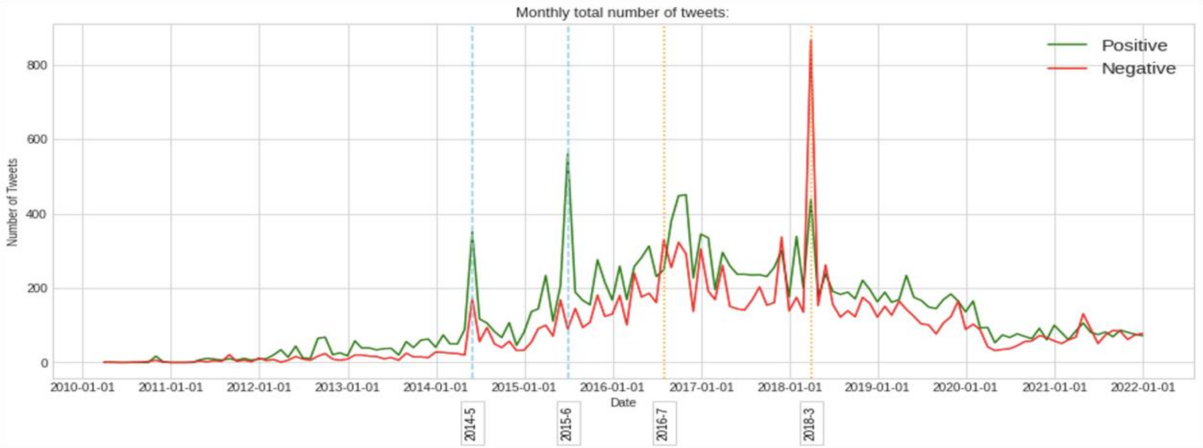


Fig: Time series analysis of positive and negative sentiments

Month	Positive sentiment		Negative sentiment		Possible causes/ Events
	All	USA	All	USA	
May 2012	✓	✓			Google revealed its AV prototype + Nevada became first state issue AV license
February 2015	✓				UK allowed AV testing
July 2016			✓	✓	Tesla autopilot crash in Florida
March 2018			✓	✓	Uber pedestrian crash in Arizona

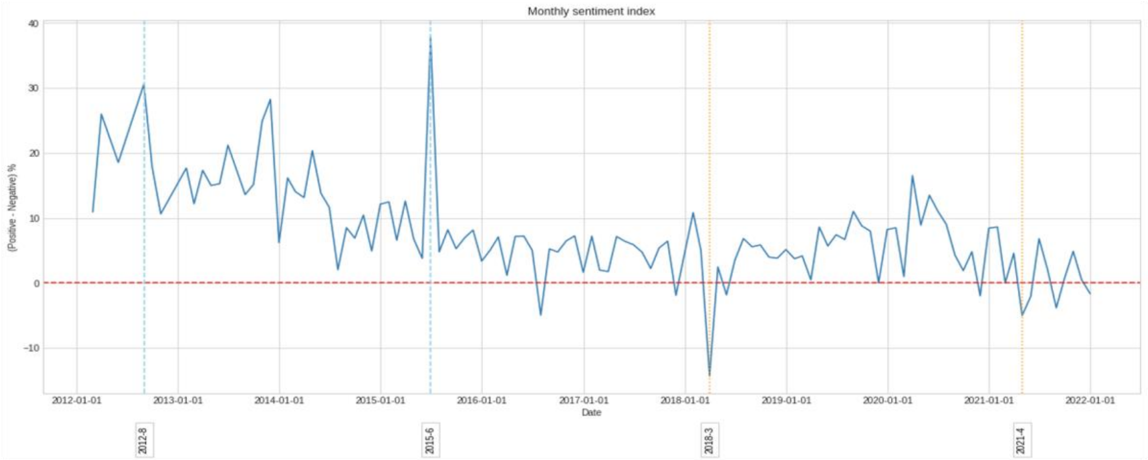
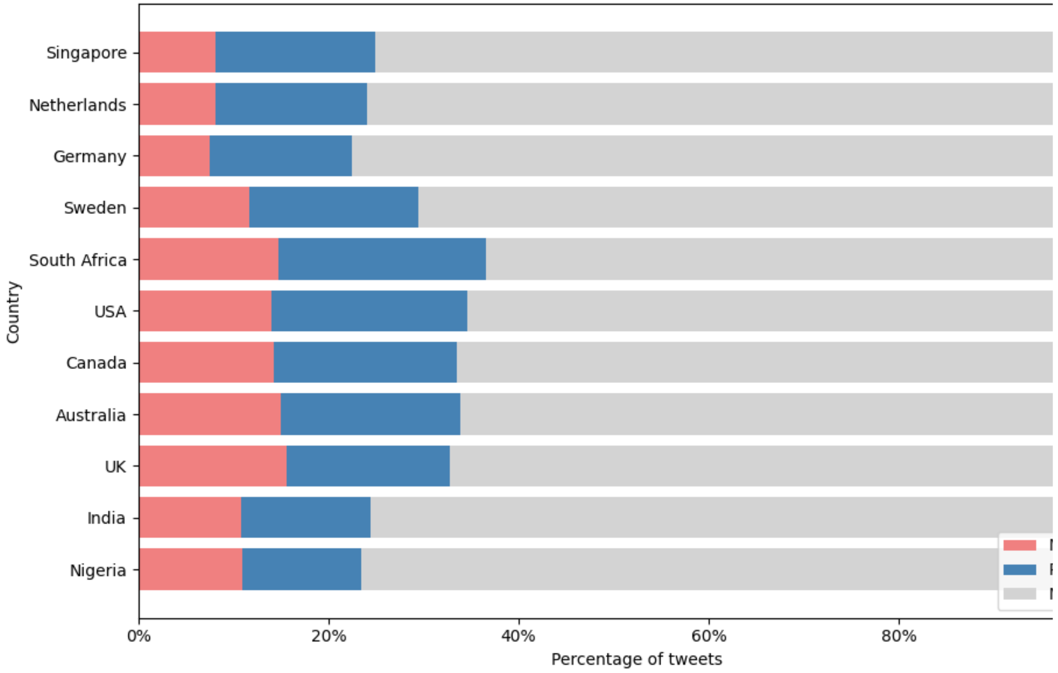
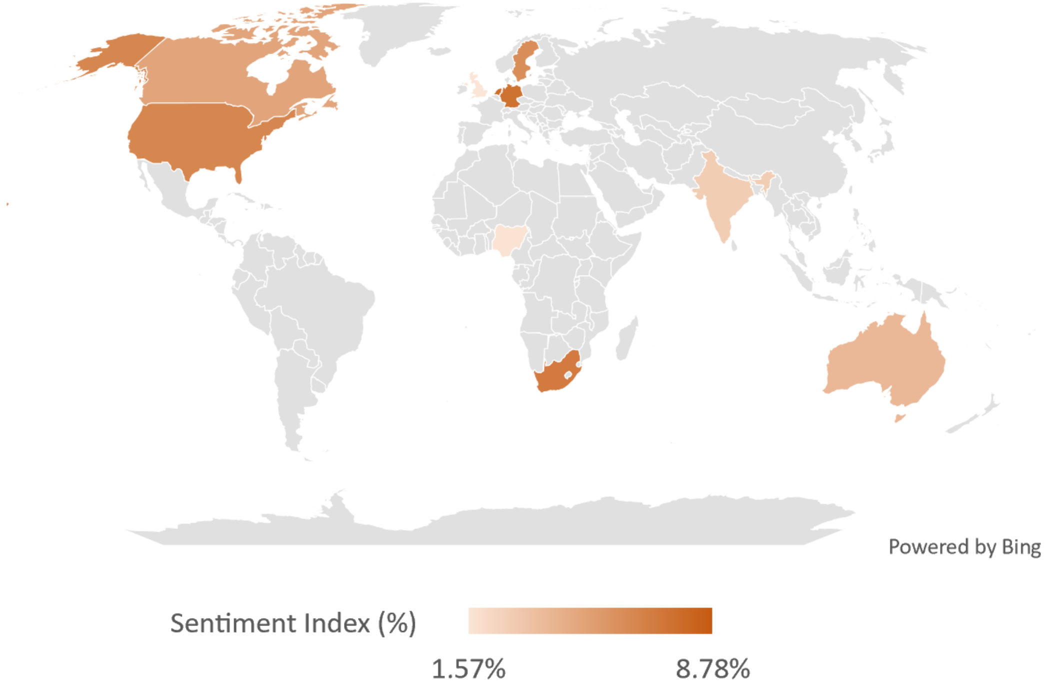


Fig: Time series analysis of sentiment index


Singapore
Netherlands
Germany
Sweden
South Africa
USA
Canada
Australia
UK
India
Nigeria
Country
0%
20%
40%
60%
80%
Percentage of tweets


Powered by Bing
Sentiment Index (%)
1.57%
8.78%

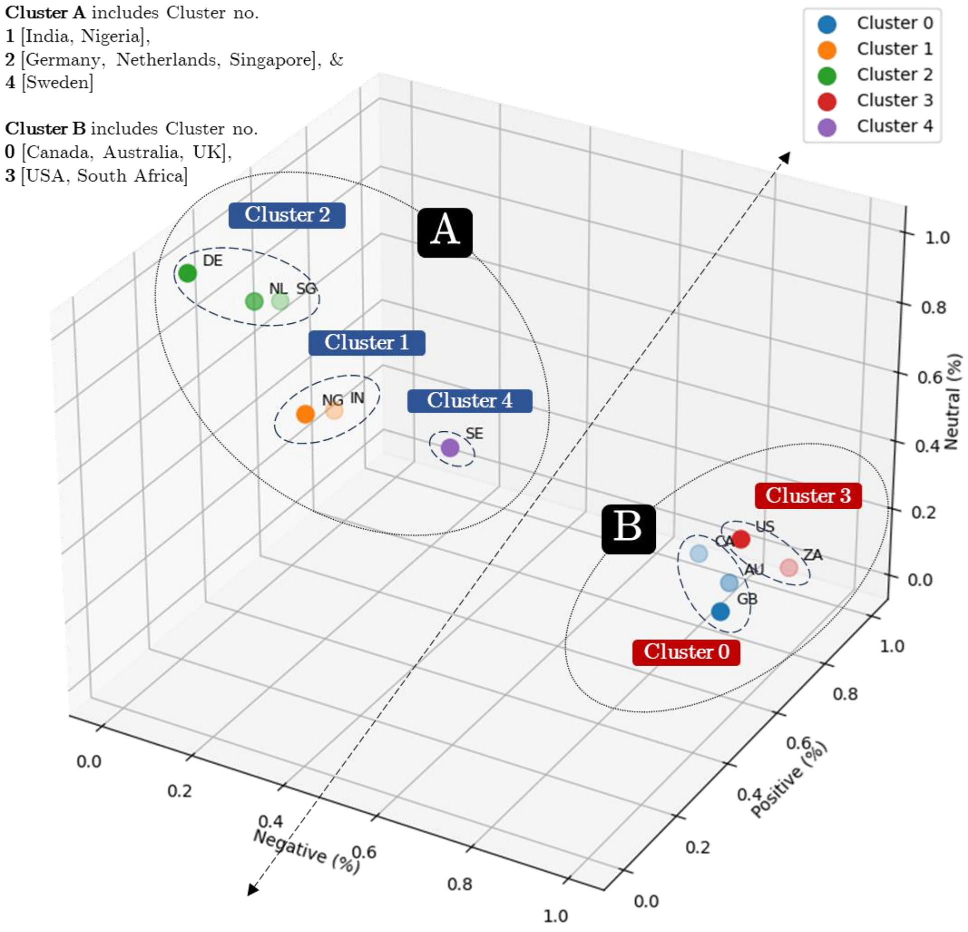
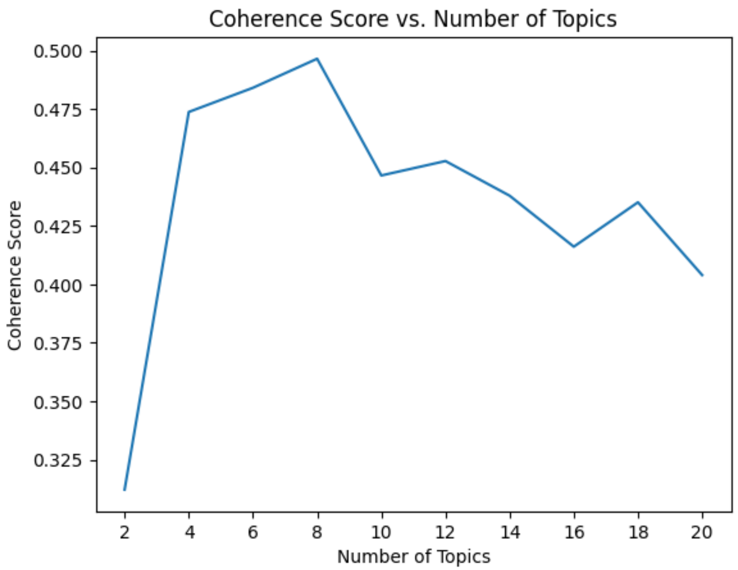


Fig: K-means clustering based on normalised polarity scores

Cluster B		Cluster A		
C-3	C-0	C-2	C-4	C-1
USA- Car dependent Uncertainty accepting	CA and AU- similar backdrop UK- safety scepticism	SE- high ICT awareness	DE- innovation capability NL- smart road infra SG- legislation	IN & NG- lower awareness + affordability
NPS: >0.8	0.5-0.7	0.5	0.2-0.5	<0.1

Identification of topic

Key themes



- Used both **Gensim and BERTopic** packages [work in progress]
- Disentangling **time and country specific effects**
- Decided **8 key topics** based on **coherence score (Gensim package)**
  - **2 major topics** and other 6 overlapping

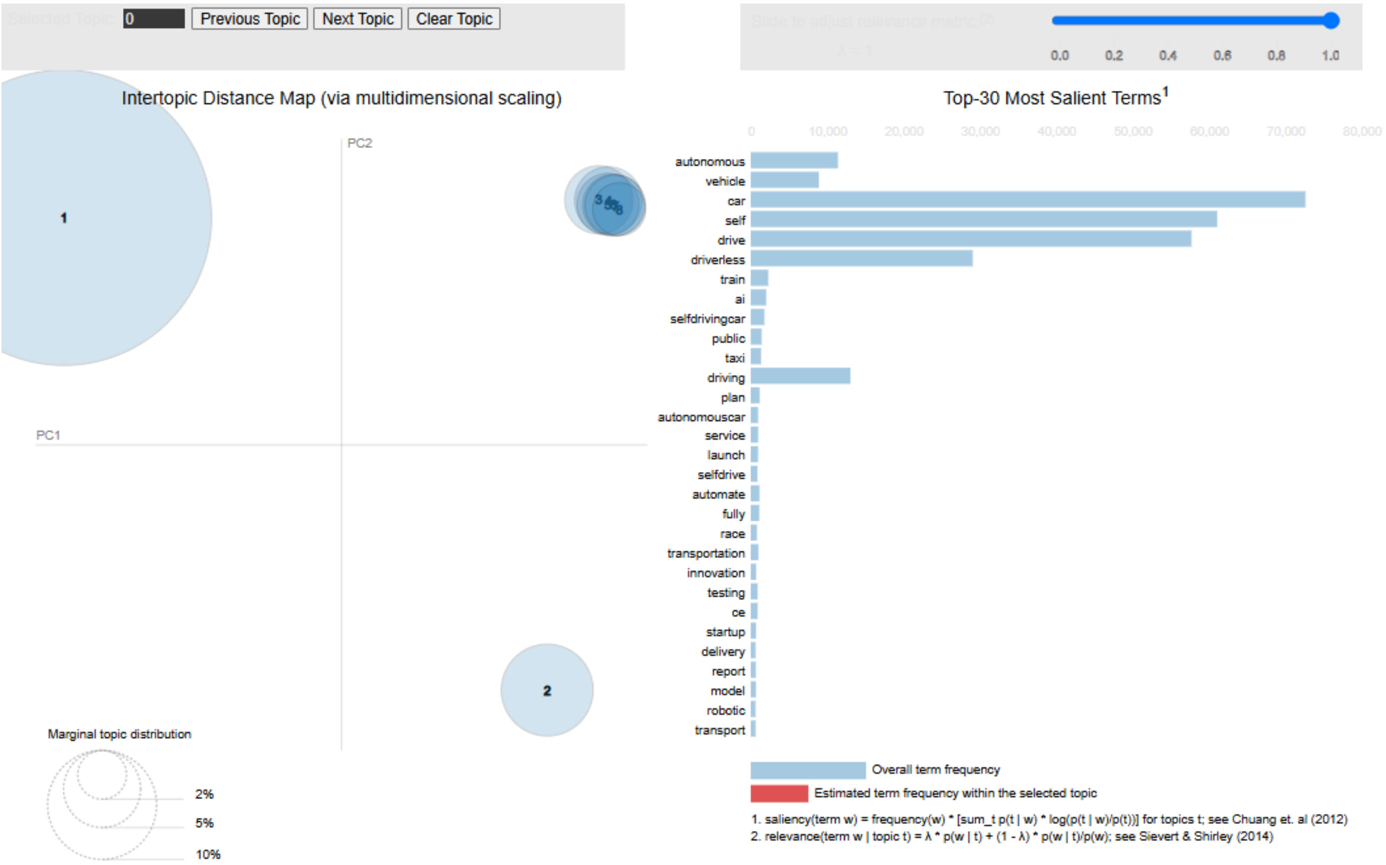


Fig: Inter-topic distance map and Keywords for each topic

- Key discussion themes can be grouped into: (1) **AV enthusiasm** and (2) **AV concern**
- **AV enthusiasm** includes-
  - Technology/ Innovation **(id: 1)**
  - Automation advantages (safety/ connectedness) **(id: 2 , 3, 8)**
  - Service types and infrastructure **(id: 3)**
- **AV concern** includes-
  - Accident **(id: 4)**
  - Ethical and moral responsibility **(id: 9)**
  - Critical decision and dilemmas **(id: 9)**

- **Please note that topic_0 is outlier** (consists of words which don't add to deciphering underlying theme)
- Topic_6 could be off topic (tweets related to driverless train services)

Documents and Topics

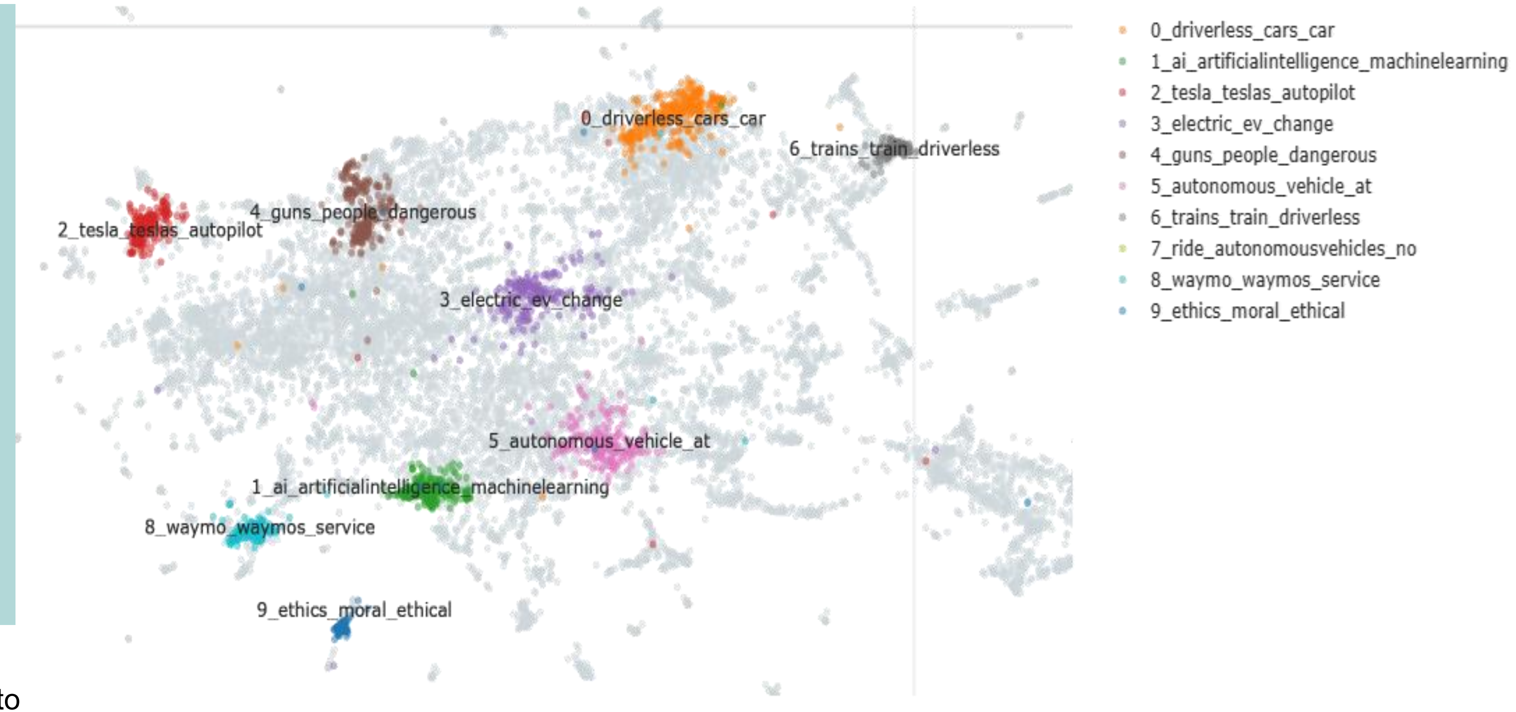


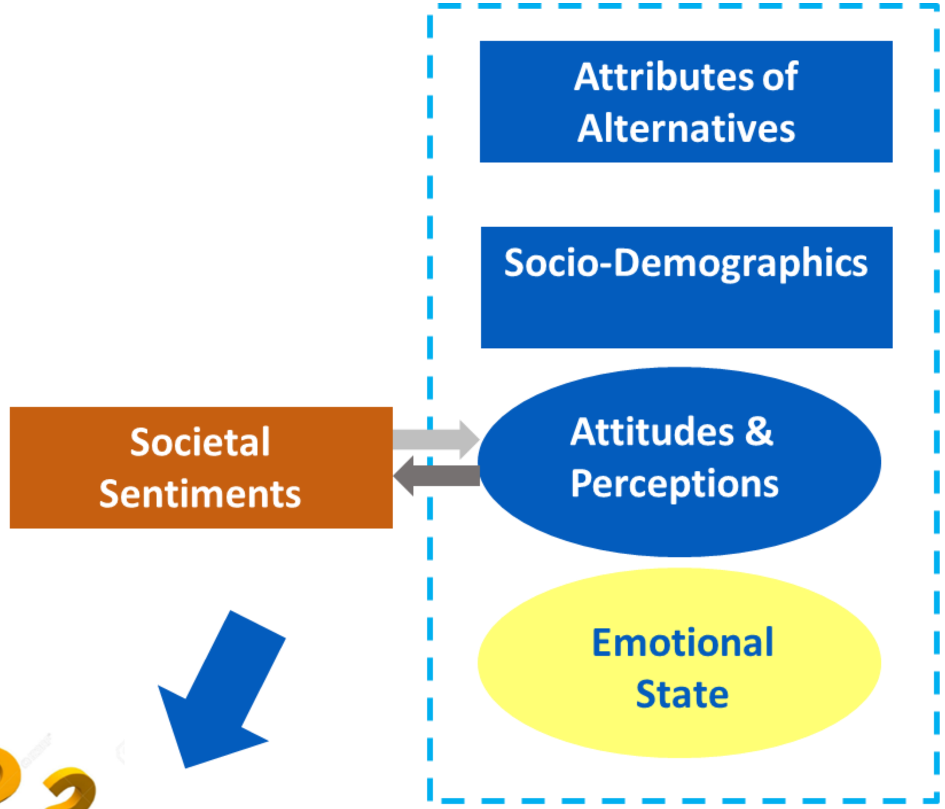
Fig: Topic visualisation (2018 tweets)

Ongoing work:

1. How have the topics evolved over the years?

   - Correlation with the stage of implementation?

2. How different are the topics for countries within the same spatial cluster?
3. Developing frameworks on how the findings can be integrated with traditional travel behaviour models



$$P(y_{in}) = f(x_{in}, z_n, \beta)$$

Questions?

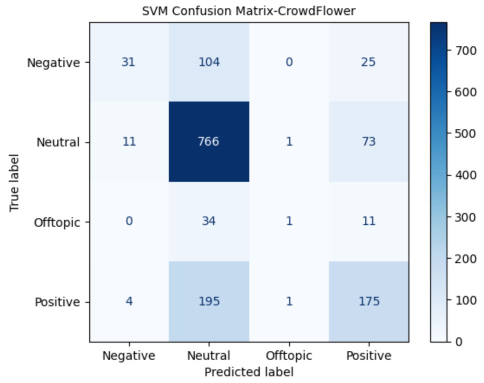
c.f.choudhury@leeds.ac.uk

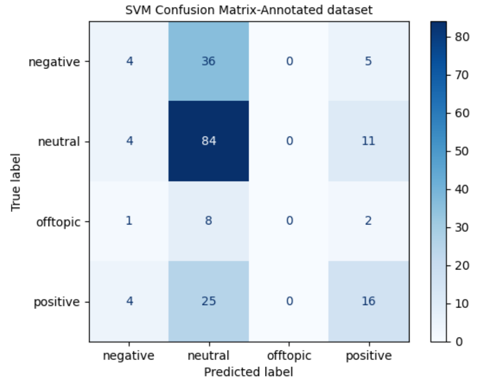
Codes: https://github.com/arashk1990

NEXt generation activity and travel behavioUr modelS: Bringing together choice modelling, data science and ubiquitous computing (MR/T020423/1)

UKRI Future Leaders Fellowships

Support Vector Machine: Best performance





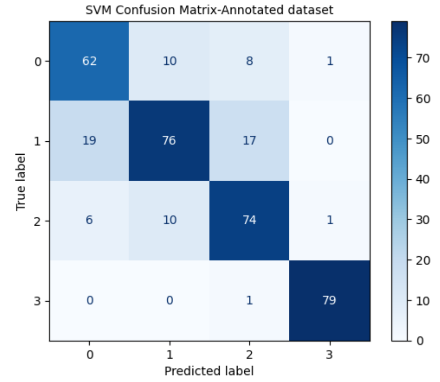


Table: Classification results of SVM

	CrowdFlower data			Annotated dataset			Augmented annotated dataset		
Index	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Negative	0.449	0.157	0.233	0.353	0.158	0.218	0.813	0.691	0.747
Neutral	0.716	0.886	0.792	0.518	0.914	0.661	0.625	0.655	0.640
Positive	0.623	0.480	0.542	0.632	0.231	0.338	0.718	0.782	0.749
Off-topic	0.333	0.029	0.054	0.000	0.000	0.000	0.965	0.976	0.971
Accuracy			**0.687**			**0.515**			**0.775**
macro avg	0.530	0.388	0.405	0.376	0.326	0.304	0.780	0.776	0.777
weighted avg	0.656	0.687	0.654	0.472	0.515	0.437	0.779	0.775	0.775