SC4021 Information Retrieval

Sentimental Analysis on Electric Vehicles (EVs)

Group 24
Ong Zhi Ying, Adrian
Takesawa Saori
Cheong Yong Wen
Kwok Zong Heng
Mandfred Leow Hong Jie
Mao Yiyun

Roles

Adrian

- Data Crawling
- Solr Indexing

Saori

- UI design
- UI implementation

Casper

- Annotation
- Vader Classification
- Textblob Classification

YiYun

- Classification
- Test set selection
- Roberta classification
- Innovation-majority voting

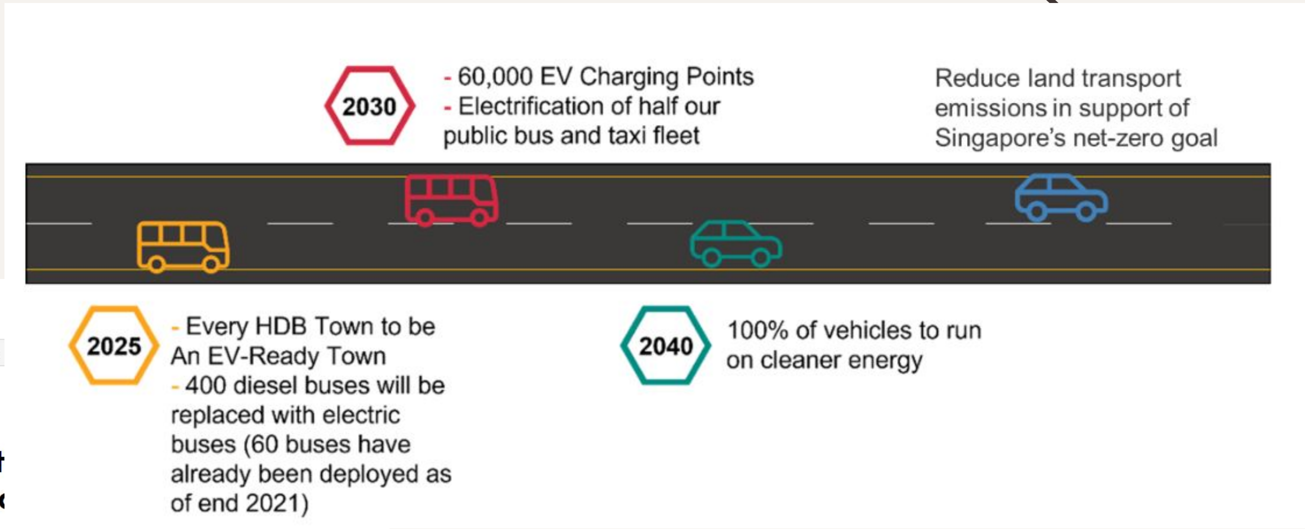
Zong Heng

- Annotation
- BERT Classification
- Innovation

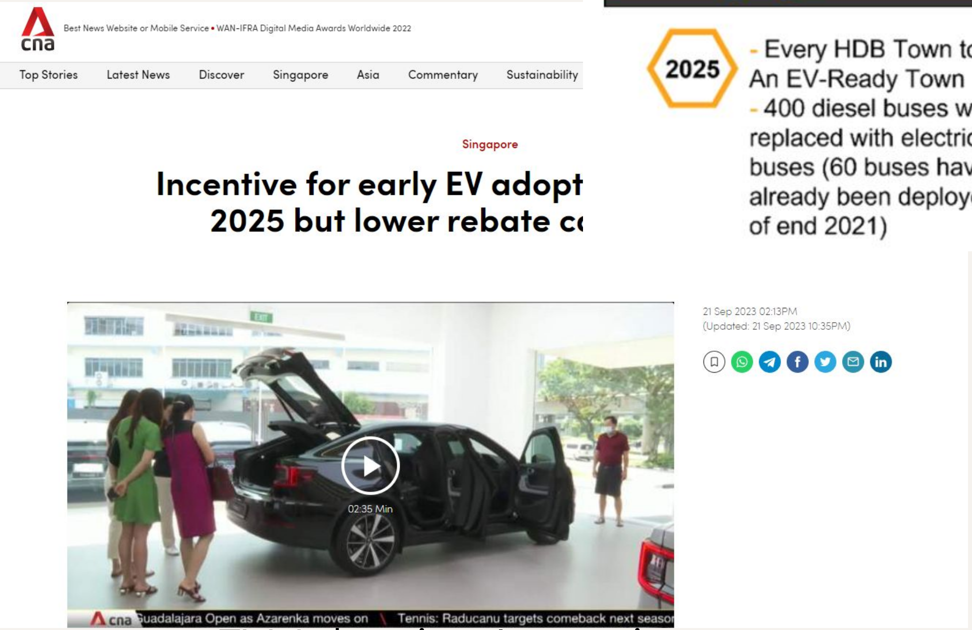
Mandfred

- Innovation

Background



LTA EV Vision



EV Adoption Incentive

- By 2040, consumers will soon be required to make a decision on which electric vehicle to purchase.

Objective & Intended Impacts

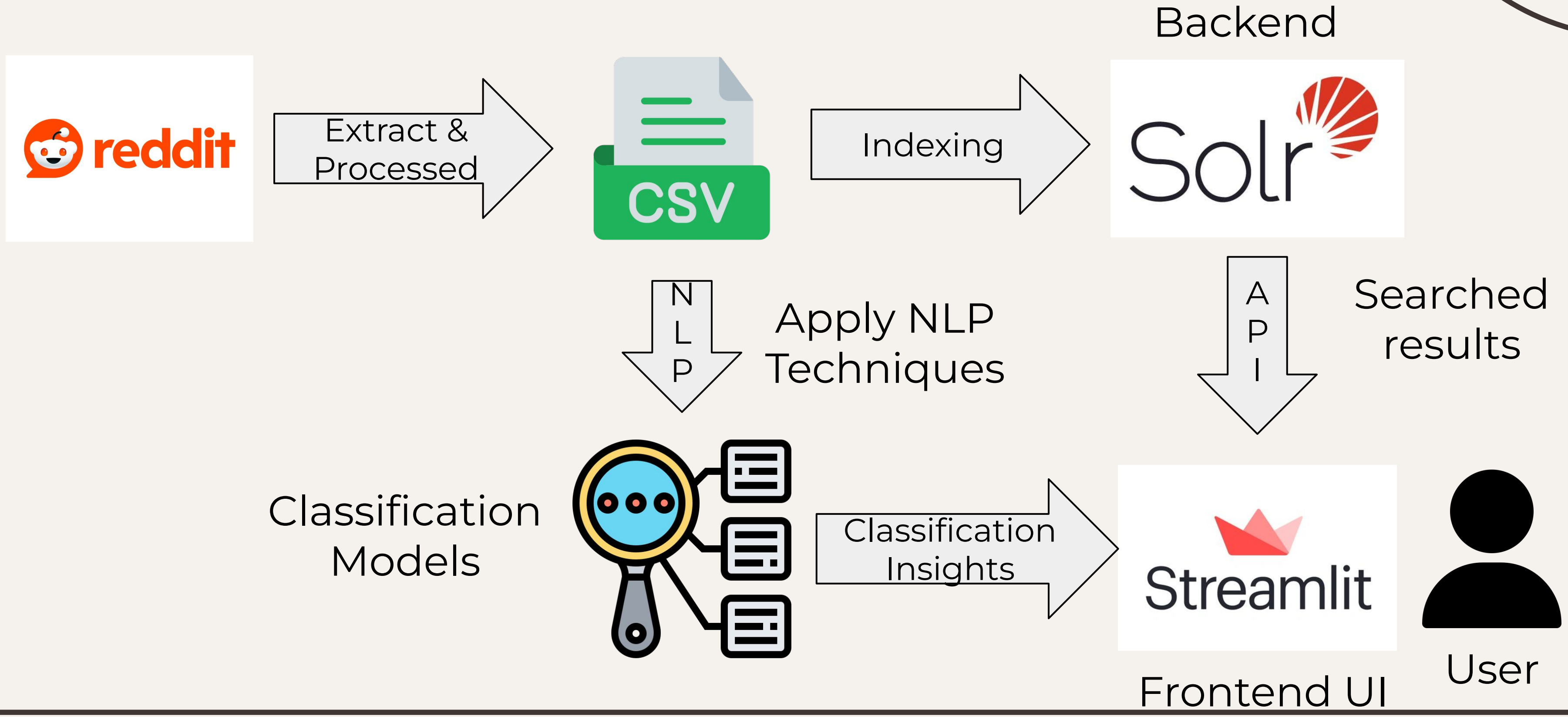
Objective

- Design a Information Retrieval System by curating, processing, analysing and presenting data and sentimental insights

Intended Impacts

- Well equipped consumers to select the best EV brand for them
    - Public sentiment of popular EV brands
    - Features of an EV brand
    - Pros and cons of each EV brand

Overview Architecture



Data Crawling

Data source

- Posts & Comments from
  - EV Brand subreddits in Reddit
  - General EV discussion subreddits

Crawling method

- Reddit PRAW
  - Extract the top 100 post per subreddit
  - Crawl all comments for those posts
- Removal of Bots, Mods posts & Comments
- Only consider top-level comments

Crawled Data statistics

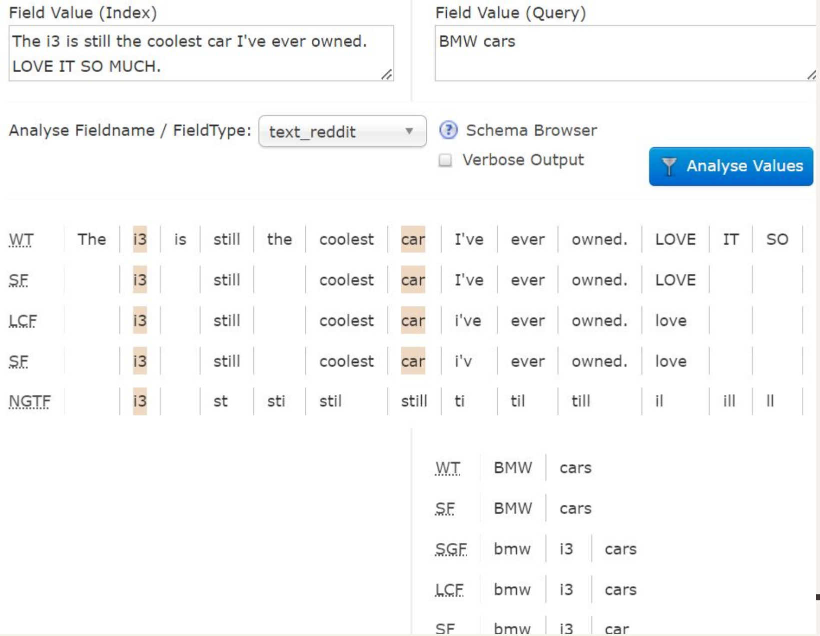
Basic Data pre-preprocessing

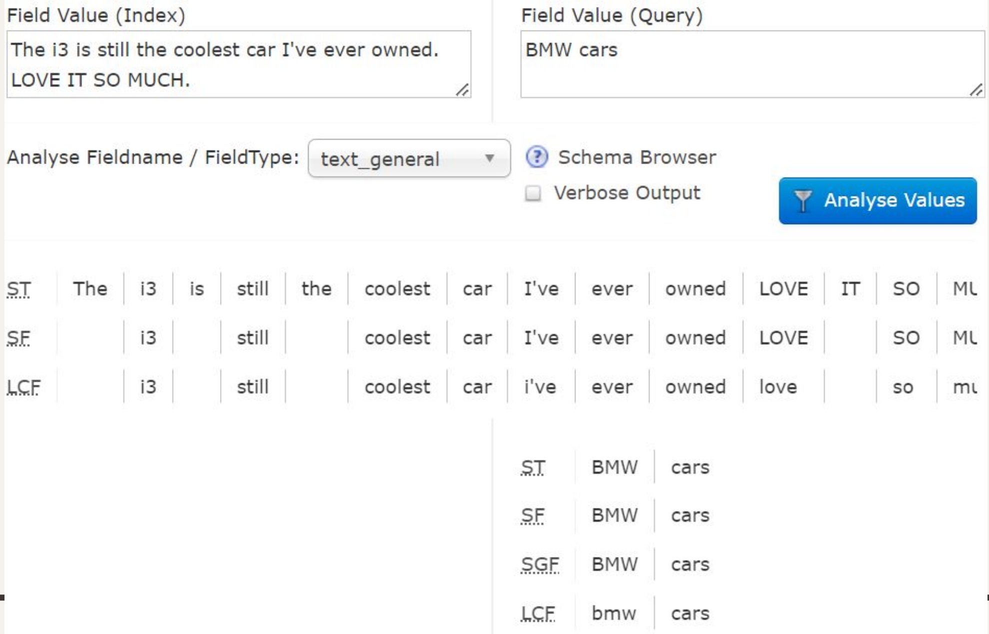
- Microtext/slang mapping (LOL -> Laugh Out Loud)
- Emoji handling (😊 -> :Smiling_face:)

Subreddits crawled	13
Number of crawled posts	1,229
Number of crawled comments	48,194
Total number of tokens in the corpus	1,176,272
Total number of unique tokens in the corpus	74,055

Data Indexing Innovations

- Spell Checking
- Custom filters
    - Synonyms for mapping model to brand names (I3 -> BMW I3)





Classification Approaches

- **VADER**
- **Textblob**
- **BERT**
- **Twitter-roBERTa-base**
- **roberta-large-mnli**

Classification

Lexicon and rule-based

VADER

- Specifically attuned to sentiments expressed in social media
- Pre-built lexicon that contains words and phrases
- Grammatical and syntactical rules

Machine Learning algorithm

Textblob

- Pre-trained on labeled dataset
- Flexibility and adaptability

Classification

RoBERTa architecture

Twitter-roBERTa-base

- Remove the NSP objective
- Dynamic masking during pre-training
- Training on a large corput
- Around 124 million tweets

RoBERTa architecture

RoBERTa-mnli

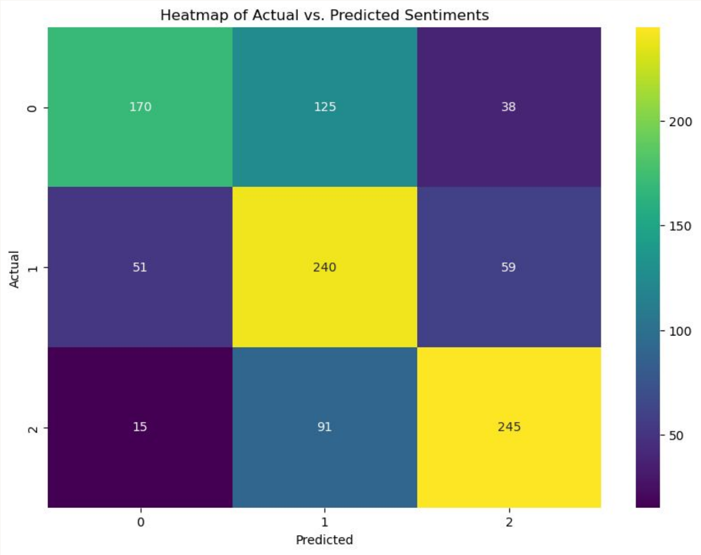
- Fine-tuned on MNLI corpus
- Exposed to various linguistic styles

Classification

Bidirectional approach

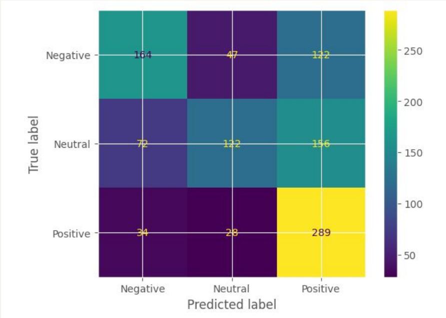
BERT

- Analyzes text by considering both left and right of every word simultaneously
- Process words in batches, enabling faster and more efficient analysis

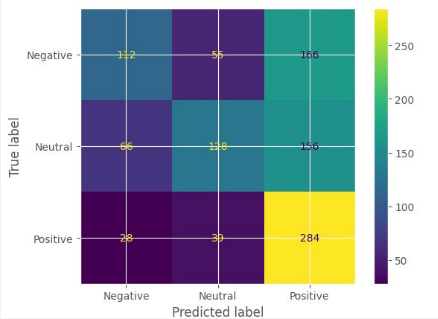


	precision	recall	f1-score	support
negative	0.72	0.51	0.60	333
neutral	0.53	0.69	0.60	350
positive	0.72	0.70	0.71	351
accuracy			0.63	1034
macro avg	0.65	0.63	0.63	1034
weighted avg	0.65	0.63	0.63	1034

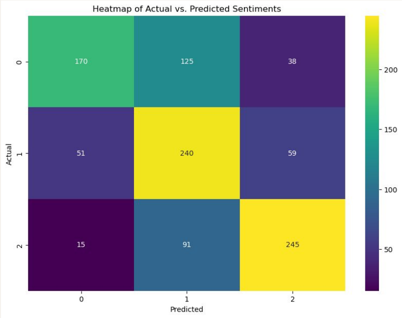
Classification Results



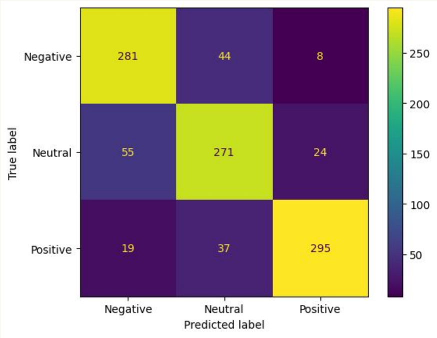
VADER



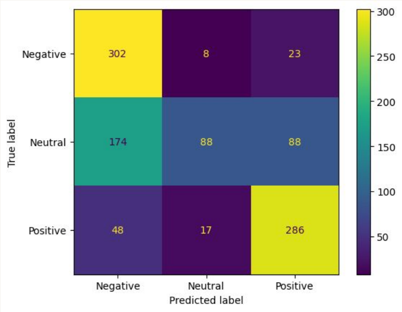
TextBlob



Bert



Twitter-roBERTa-base



roBERTa-mnli

Bright yellow means more true positives

Innovation (Stack Ensemble)

BERT, Logistic Regression and Random Forest

- Split the annotated data of train to test at a ratio of 75/25
- Fine Tuning BERT pre-train model with own dataset
- Integrate BERT model prediction with Logistic Regression and Random Forest
- Predictions of BERT,Logistic Regression and Random Forest were used as input feature for logistic regression model
- Trained on combined prediction to learn final judgements on the sentiments.

	precision	recall	f1-score	support
negative	0.83	0.58	0.68	83
neutral	0.65	0.77	0.70	94
positive	0.74	0.81	0.77	79
accuracy			0.72	256
macro avg	0.74	0.72	0.72	256
weighted avg	0.73	0.72	0.72	256

Bert Model only

	precision	recall	f1-score	support
negative	0.76	0.63	0.69	83
neutral	0.65	0.73	0.69	94
positive	0.78	0.81	0.80	79
accuracy			0.72	256
macro avg	0.73	0.72	0.72	256
weighted avg	0.73	0.72	0.72	256

Stacked

Innovation (Voting Ensemble)

VADER, BERT, and roBERTa-MNLI for majority voting

```
Accuracy: 0.5560928433268859
Precision: 0.5560928433268859
Recall: 0.5560928433268859
F1 Score: 0.5560928433268859
```

VADER

```
Accuracy: 0.6334622823984526
Precision: 0.6334622823984526
Recall: 0.6334622823984526
F1 Score: 0.6334622823984526
```

BERT

```
Accuracy: 0.6537717601547389
Precision: 0.6937598452290512
Recall: 0.6537717601547389
F1 Score: 0.6537717601547389
```

roBERTa-mnli

```
Accuracy: 0.6760154738878144
Precision: 0.6760154738878144
Recall: 0.6760154738878144
F1 Score: 0.6760154738878144
```

Majority Voting Model