ML Project WF

27 May 2022 09:32

Eg: Control center Dashboard
← news feeds
↖ news feeds
manually classifying
→ irrelevant
→ natural emergency
→ traffic incident
→ social unrest
→ Criminal activity
Accenture facilities

1. → Understanding the business problem
   → Classification of incidents from news feed (high volume)
   Cost of man power required.

2. → Feasibility testing.
   → has this problem been solved with ML in past.
   → Literature Review → OK

3. → Defining success metrics → based on HLP or SOTA → 95%.
   → human level performance → 94%
   Classification accurately → 95% or 5% error rate.
   precision
   recall/f1

4. → Identifying data requirements
   → Training data → Already classified news feed incidents. | supervised

5. → Collecting training data
   → Proprietary customer data ✓
   → Open source data sets ✓
   → Web Scraping ✓
   → manually labelled ✓
   → Historically labelled data
   → (+) → Training data

→ Cleaning and preprocessing data
numerical data
→ Imputing missing values
→ Fixing data imbalance
→ Normalizing data
→ Dimensionality Reduction.
→ removing empty/irr examples
→ irrelevant: 90% remains all classes 10%

ht	Area
5	1000
6	2000
	3000

normalize

PCA TSNE
→ Train Test split

10,000 → 9000 irr → ROSE → Random over sampling of minority data
100 → 1
2000 ← 100 - Acc
2000 ← 300 - crime
2000 ← 800 - Social unrest
2000

→ Creating a baseline model
50K → training data → train 80%
→ validation 20%
[8K] [2K]
(40K) baseline → 80% accuracy → 20% error →
(10K)
2K error records → error analysis

(10k)

2k error records → error analysis

→ One particular class has low accuracy.

↓

→ Improving model performance (based on error analysis & data driven)

↳ ✓95% acc || pretrained model → finetune.

→ bert model → Wikipedia corpus for language modelling. [predicting the next word]

→ Model deployment
→ Monitoring model o/p for drift.

final prod. model

@ deployment → 24 hours news feed

80% - irre → ±10%
10% - traffic → ±10% → alert.
5% - rain
3% - climate
etc.

---

(1) Understand the business usecase:-

Figure out the client requirement → eg: Automate Disaster mgmt team's manual classification of news feed in different categories eg: irr. climate or ins, etc.

(2) Figure out the feasibility:

literature Review

(3) Success metric based on literature review results

✓SOTA models
✓HLP

→ Classification
→ Acc. ⇒ 95%
→ Recall
→ Precision

(4) Identify data requirement & data sources.

4. Identify data requirement & ... sources.
   - Classified News feed.
   - ✓ Use historic data available in the system.
   - Web Scrape
   - Open source dataset.
   - Manually label the data.

6. Data Cleaning
   - txt
     - Remove empty values.
     - Balancing the dataset.
       - ROSE
       - SMOTE
   - numerical
     - data normalization
     - dimensionality reduction → PCA / ↘ TSNE
     - Imputing empty values. (mean/median/mode)
       - Fill
       - mean → outlier sensitive
       - median → less affected by outliers
       - mode → categorical → Freq.

7. Baseline Model.
   - ↳ 80% acc
   - 20% error → error analysis in data & model.

8. Model Improvements
   - After the success metric is met.

9. Model Deployment

10. Model maintainance → Output distribution of labels monitor karte hain.