Using natural language processing to classify sentences in RCT’s

Megan Pete

Introduction

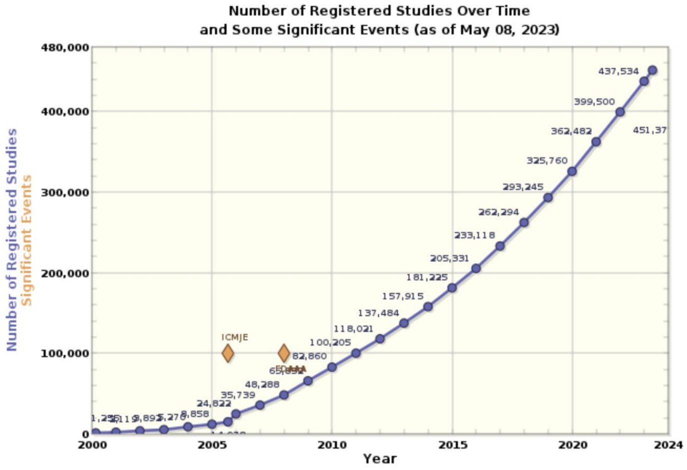
Introduction

Academic papers:

- The number of scientific studies steadily **increasing** over time since the year 2000
- High volume of new research being published → necessary for researchers to **efficiently** sift through articles
- Whether or not a paper is of use to a researcher may depend on **various factors**, such as whether it includes particular methods, results or topics

Role of NLP:

- **NLP:** Used to analyse and understand the use of languages in text
- **Tasks**: Sentiment analysis, translation, classification, summarising etc.
- NLP may be able to help researchers to quickly and efficiently skim abstracts of studies in order to identify useful papers



Source: https://ClinicalTrials.gov

Pubmed 200k dataset

Description

- Dataset based on PubMed for sequential sentence classification
- Contains 200,000 abstracts from RCT’s
- Each sentence is labelled as: **Background, objective, method, result or conclusion**

Purpose

- Generally: provide large dataset for sequential short text classificaiton
- Specifically: help researchers skim abstracts more efficiently

Research aims

Aim 1:

- Predict the category of sentences in abstracts of randomised controlled trials

Aim 2:

- Identify clusters of words that describe each main sentence category

Methods

1. Pre-processing
2. Supervised methods
3. Clustering

Data pre-processing

1. **Tokenize each abstract**
- Split data into labelled sentences

2. **Clean sentences**
- Remove stop words (eg. ‘and’, ‘you’, ‘very’, ‘more’ etc.)
- Remove punctuation, capital letters, symbols, numbers
- Stemming: Reduce words into their base form (eg. running → run)

3. **Create bag of words model**
- Vectorize sentences using *TfidfVectorizer* function (Term-frequency inverse-document-frequency vectorizer)
- Creates matrix: Columns represent dictionary of words; rows represent TF-IDF values (measure of importance of each word in the sentence)
- Term frequency: how often a word shows up in a sentence
- Inverse document frequency: penalises words that show up in many sentences

4. **Train-test split** (80:20)

Supervised methods

- Random forest and logistic regression
- Parameter tuning using 5-fold cross-validation and grid search
- Upsampling
- Metrics:
  - Precision, recall, F1 score, AUC
- Feature importance:
  - **Logistic:** Absolute values of coefficients
  - **RF:** Based on decrease in impurity
    (Greatest decrease in impurity = most important feature)

Clustering

- K-means: computational efficiency and interpretability
- Choosing number of clusters: Silhouette scores and SSE
- Visualise results: heatmap, t-SNE plot, PCA plot

Results

- Descriptive statistics
- Aim 1 (supervised)
- Aim 2 (unsupervised)

Training set description

Training set frequencies:

```
RESULTS: 10186
METHODS: 10546
OBJECTIVE: 2514
BACKGROUND: 3908
CONCLUSIONS: 4846
```

Proportion of each part of speech per sentence type

label	Nouns	Verbs	Adjectives	Adverbs
BACKGROUND	55.36	15.74	24.98	3.92
CONCLUSIONS	54.86	16.30	24.42	4.42
METHODS	58.02	17.06	21.95	2.97
OBJECTIVE	58.96	13.93	24.47	2.64
RESULTS	60.17	13.66	21.21	4.96

Random forest: results

Test set:

	Class	Precision	Recall	F1-score	AUC
0	BACKGROUND	0.60	0.34	0.44	0.88
1	CONCLUSIONS	0.61	0.46	0.53	0.88
2	METHODS	0.71	0.87	0.78	0.93
3	OBJECTIVE	0.62	0.47	0.54	0.90
4	RESULTS	0.75	0.81	0.78	0.92

Training set:

	Class	Precision	Recall	F1-score	AUC
0	BACKGROUND	0.96	0.87	0.91	1.00
1	CONCLUSIONS	0.97	0.90	0.93	0.99
2	METHODS	0.88	0.98	0.93	0.99
3	OBJECTIVE	1.00	0.87	0.93	0.99
4	RESULTS	0.96	0.95	0.96	0.99

[Optimal parameters: Number of estimators = 300, Min samples to split = 10, Max depth = 100]

- AUC significantly higher on test set than F1 scores (F1 score more sensitive to class imbalance and overfitting than AUC)
- Best F1 and AUC scores: **Methods** and **results**

<u>Upsampling:</u>

- Upsampling **did not** improve F1 scores (increased recall but decreased precision, due to overfitting upsampled classes)
- **Improvements:** adjust the decision threshold, and tune the resampling ratio

Logistic: results

Logistic Regression Test set:

	Class	Precision	Recall	F1-score	AUC
0	BACKGROUND	0.52	0.50	0.51	0.89
1	CONCLUSIONS	0.60	0.55	0.57	0.89
2	METHODS	0.80	0.83	0.82	0.94
3	OBJECTIVE	0.55	0.49	0.52	0.91
4	RESULTS	0.77	0.80	0.79	0.93

Logistic Regression Training set:

	Class	Precision	Recall	F1-score	AUC
0	BACKGROUND	0.60	0.58	0.59	0.92
1	CONCLUSIONS	0.66	0.61	0.63	0.92
2	METHODS	0.82	0.87	0.84	0.95
3	OBJECTIVE	0.66	0.53	0.59	0.94
4	RESULTS	0.81	0.83	0.82	0.95

[Optimal parameters: C = 10, Penalty = L2]

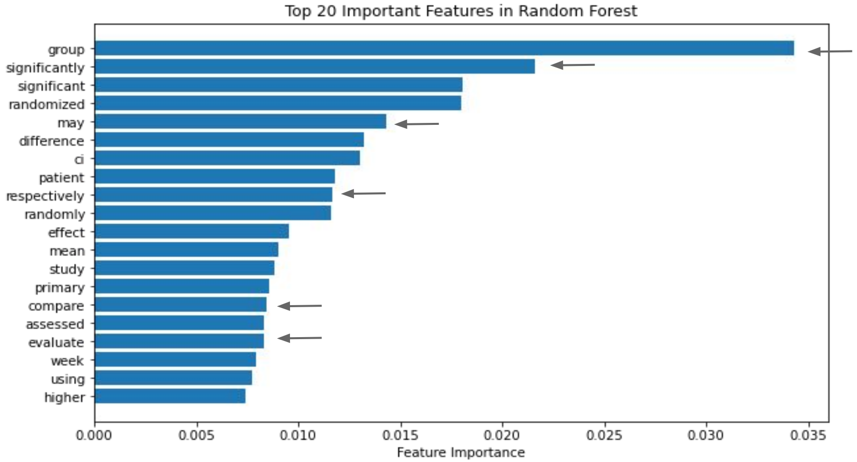
Logistic test set results are similar to RF:

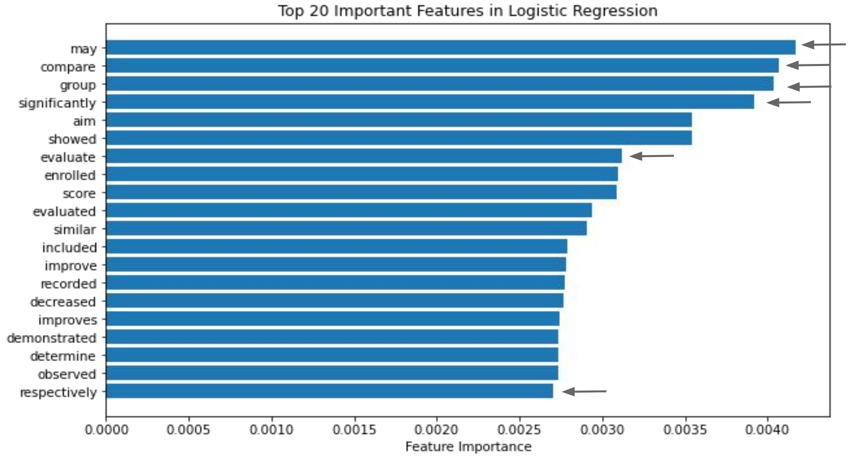
- Similar F1 and AUC scores
- AUC significantly higher than F1 scores
- Best F1 and AUC scores: Methods and results
- Upsampling objectives and background did not improve results

However:

- F1 scores on training set lower than RF, indicating **less overfitting**
- Possible explanation: RF more complex model, and does not have built in regularisation

Feature importance

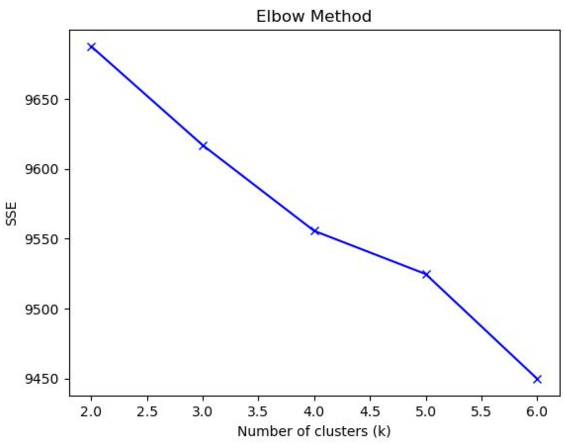


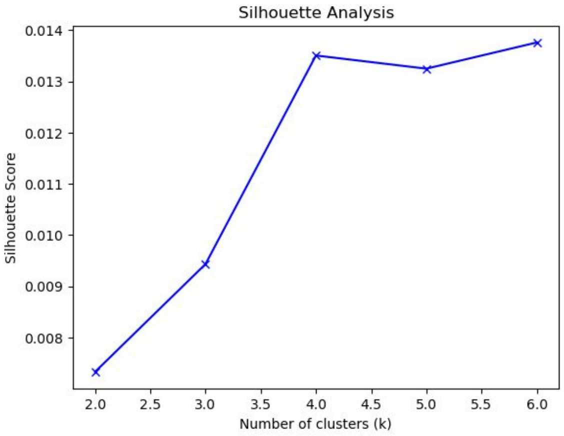


- 6 of the top 20 features in common
- Although prediction is relatively similar, different features are driving the predictions
- Feature importance: small values (distributed evenly)

K-means clustering

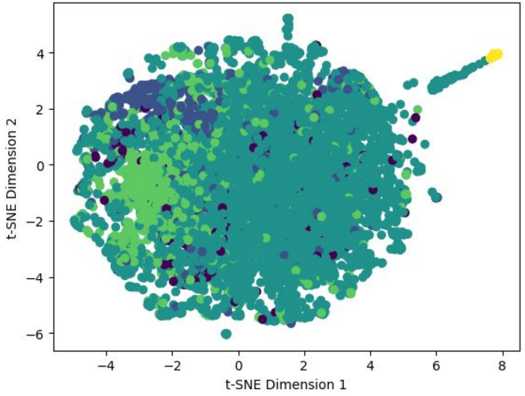
Choosing K:

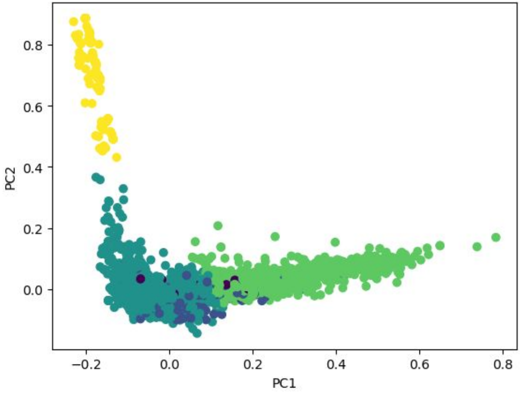




- SSE: no clear elbow
- Silhouette scores: not much change from k=4 onwards
- All values relatively close together
- Hence, prioritise interpretability and choose k=5 as there are 5 labels

K-means clustering





PCA

- Identify global patterns in the data
- Linear technique
- We see relatively separable clusters

t-SNE

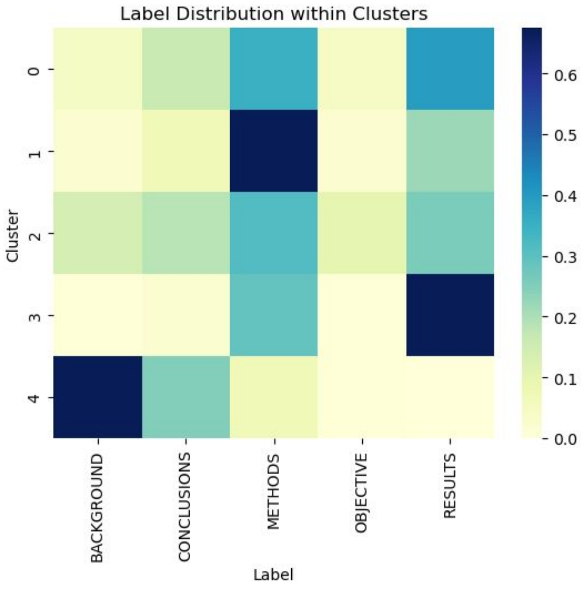
- Captures more subtle patterns
- More sensitive to outliers
- t-SNE clusters do not look separable
- Changing perplexity of t-SNE did not have large effect

Possible explanation

- Data has linear structure: PCA deals with this well
- t-SNE may focus too much on local relationships to capture global structure

K-means clustering

Interpret clusters:



- Clusters 1, 3 and 4 each correspond well to methods, results and background respectively
- Objective and conclusions do not have clear corresponding clusters
- This may be partially explained by the fact that the data is imbalanced

Conclusion

Conclusion

Supervised:

- Predicted types of sentences from RCT abstracts
- **Similar** results for both RF and logistic models, but less overfitting in logistic (possibly due to regularisation used)
- **Methods** and **results** were predicted the best
- Good AUC scores; mediocre F1 scores (class imbalance)
- Basic upsampling did not improve F1 scores (tradeoff between recall and precision)
- <u>Feature importance:</u> Small feature importance across multiple variables (many small effects rather than few large effects)
- 6 of the 20 most important features common across both models

Unsupervised:

- K-means used to group sentences into 5 clusters
- PCA showed separable clusters, but t-SNE did not, suggesting strong linear global structure
- Methods, results and background were well described by clusters which supports our supervised method findings

Conclusion

Possible improvements:

- Use full 200k dataset to improve prediction
- Work on improving F1 score for background and objectives labels by upsampling. Adjust the decision threshold, and tune the resampling ratio
- Use larger amount of words in the model (>1,000) since feature importance is somewhat evenly distributed

Questions?