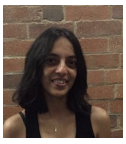
Gradient boosted model for case-ascertainment from veterinary records



Uttara Kennedy
uttara.kennedy@uq.edu.au

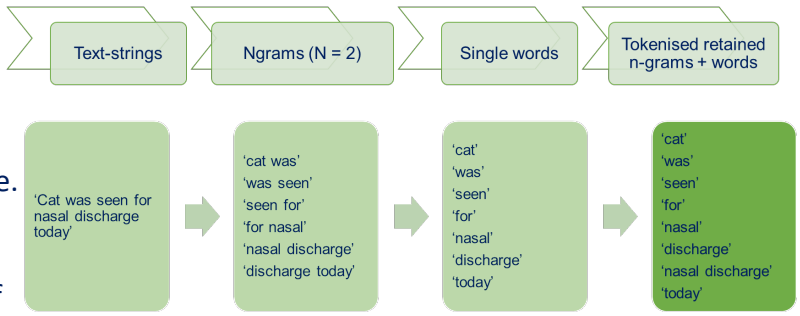
BACKGROUND:

Since veterinary records are not consistently structured, it is a challenge to get case ascertainment from large datasets spanning long time-frames. Without this information, we cannot determine prevalence and incidence of disease, or conduct risk factor analyses.

METHODS:

1. Convert records into a continuous text-string.
2. Convert text string into tokens of 2-word combinations (n-grams) and single words
3. Use tokens to train a gradient boosted model with a manually annotated sample.
4. Test the model with an out-of-sample novel dataset.
5. Obtain case ascertainment (probability of infection) for entire dataset.

Step 2: Tokenisation:



We used **machine learning** for case-ascertainment of "cat flu" from 8 years of electronic clinical records.

This enables prevalence estimates and risk factor analysis for the first time from RSPCA Qld data.

Take a picture to access publication

RSPCA Queensland

The University Of Queensland AUSTRALIA

Dr. Uttara Kennedy
Dr. Nicholas Clark
Dr. Mandy Paterson
Prof. Ricardo Magalhaes

RESULTS:

- The trained GBM was deployed on a dataset comprising 60,258 entries.
- The GBM used 1250 unique words and n-grams as predictors.
- Predicted probability had an **accuracy** of **0.95 (95% CI 0.92, 0.97)** and **F1 score 0.96**.

	True Negative	True Positive
Predicted Negative	347	12
Predicted Positive	14	118

- The most influential words correspond with domain expertise for this disease:

Word	Relative Influence	Number of occurrences in dataset
'doxycycline'	35.76	22,957
'flu'	16.87	58,858
'sneezing'	10.04	27,563
'doxybrom'	3.28	1,950
'ocular'	2.36	18,714
'steam'	1.84	10,687
'chloramphenicol'	1.3	7,379
'bid'	1.2	22,978
'nasal'	1.11	30,287
'cat'	0.87	85,533

CONCLUSION:

- Prevalence estimates and risk factor analysis can only be done if we have case ascertainment.
- Machine learning can be used on large unstructured datasets to get case ascertainment.
- The tool can be extended to other conditions (snake bites, tick paralysis and parvovirus)