# Course 4- Practical Application Project

# Text processing technique and key metrics

- Technique: I will be using the ‘Keyword windows’ technique.
- Note type: The type of notes that have been provided for mining, ‘History and Physical’ are optimal to look for complications of a disease.
- Initial keyword/s: Diabetes/ Diabetic

This is the initial keyword/s I have chosen to identify Diabetic patients with Neuropathy, Nephropathy and/or Retinopathy, after manually reviewing a few notes.

- Initial window size: 60 words (30 before and 30 after the initial keyword/s)

Since I am looking for symptoms/ complications of the disease, I will be using a rather large window. This should help capture cases where the relevant information is 1 or 2 sentences away, as is usually the case when describing the symptoms/ complications of a disease.

# Regular expression

"(?<![a-zA-Z])diabet(es|ic)?(?![a-zA-z])“

This regular expression will identify the keywords, ‘diabetes’ or ‘diabetic’ by identifying the root word, ‘diabet’ followed by ‘es’ or (as indicated by ‘|’) ‘ic’. The ‘?’ at the end of the keyword specifies that the regex be matched if there is 0 or 1 occurrence of the target word. The keywords are also preceded and followed by look ahead and look behind groups. The look behind group, ‘?<![a-zA-Z]’ translates to- Only match the keywords if they are not preceded by upper- and lower-case A to Z. Similarly, the look ahead group, ‘?![a-zA-Z]’ translates to- Only match the keywords if they are not followed by upper- and lower-case A to Z.

In addition, the ‘ignore_case’ flag will be enabled, to make the regex case- insensitive.

# Confusion matrix: Any one of three complications

| | | ANY_DIABETIC_COMPLICATION | |
|---|---|---|---|
| | | 1 | 0 |
| ANY_COMPLICATION | 1 | 19 | 4 |
| | 0 | 7 | 81 |

Sensitivity : 0.73

Specificity : 0.95

Pos Pred Value : 0.83

Neg Pred Value : 0.92

# Confusion matrices for each of the three complications

| | | DIABETIC_NEUROPATHY | |
|---|---|---|---|
| | | 1 | 0 |
| neuropathy | 1 | 12 | 2 |
| | 0 | 3 | 94 |

Sensitivity : 0.80
Specificity : 0.98
Pos Pred Value : 0.86
Neg Pred Value : 0.97

| | | DIABETIC_NEPHROPATHY | |
|---|---|---|---|
| | | 1 | 0 |
| nephropathy | 1 | 6 | 1 |
| | 0 | 4 | 100 |

Sensitivity : 0.60
Specificity : 0.99
Pos Pred Value : 0.86
Neg Pred Value : 0.96

| | | DIABETIC_RETINOPATHY | |
|---|---|---|---|
| | | 1 | 0 |
| retinopathy | 1 | 1 | 2 |
| | 0 | 1 | 107 |

Sensitivity : 0.50
Specificity : 0.98
Pos Pred Value : 0.33
Neg Pred Value : 0.99

# Case count

**Any diabetic complication**

- **Correctly identified: 100**
  - **Positive: 19**
  - **Negative: 81**
- **Incorrectly identified: 11**
  - **False positives: 4**
  - **False negatives: 7**

Neuropathy

- Correctly identified: 106
  - Positive: 12
  - Negative: 94
- Incorrectly identified: 5
  - False positives: 2
  - False negatives: 3

Nephropathy

- Correctly identified: 106
  - Positive: 6
  - Negative: 100
- Incorrectly identified: 5
  - False positives: 1
  - False negatives: 4

Retinopathy

- Correctly identified: 108
  - Positive: 1
  - Negative: 107
- Incorrectly identified: 3
  - False positives: 2
  - False negatives: 1

# Examples of correctly identified notes and reason for correct identification

- Note ID 4 is correctly identified as negative for any diabetic complication because though the TEXT column contains the keyword, ‘diabetes’, the keywords identifying complications, ‘neuropathy/ neuropathic, nephropathy/ nephropathic’ and/or ‘retinopathy/ retinopathic’ are absent.
- Note ID 6 is correctly identified as positive for any diabetic complication because it contains two keywords required to identify a case as positive for diabetic complications, ‘diabetic’ and ‘nephropathy’.

# Examples of incorrectly identified notes, reason for incorrect identification and alternative approaches

False positives for any diabetic complication

- Note ID 1 is false positive because the window is too narrow. The keywords, ‘family history’ that could’ve identified that the keywords, ‘diabetic’ and ‘retinopathy’ refer to the patient’s family and not the patient, fall outside the 60-word window and cannot help exclude it. Alternative approach: Make window broader (but is it worth it for 1 sample out of 111?)
- Note ID 21. Contains the keywords, ‘diabetes’ and ‘neuropathy’ but the word ‘neuropathy’ has been used in the context that the patient denies having ‘neuropathy’ and thus, has no complications, though the patient is diabetic. Alternative approach: Use ‘denies’ as keyword to exclude case (but there is a possibility of picking up false negatives).
- Note ID 41. Reason for false positive unknown.
- Note ID 136. Contains keywords, ‘diabetic’ and ‘retinopathy’ but in the phrase, ‘does not show any evidence of diabetic retinopathy’ i.e the patient does not have any diabetic complications but has been incorrectly identified as positive due to the presence of the keywords. Alternative approach: Use ‘does not show any evidence’ as keyword to exclude case.

# Examples of incorrectly identified notes, reason for incorrect identification and alternative approaches

False negatives for any diabetic complication

- Note ID 7 is false negative because the keyword ‘Neuropathy’ falls outside the window of 30 words behind or ahead of the keywords, ‘diabetes’ and ‘diabetic’ that are present in the TEXT column. Alternative approach: Make window broader.
- Note ID 13. Cause of false negative unknown.
- Note ID 14 is false negative because of a typo. ‘Neuropathy’ has been spelt as ‘neuopathy’. Alternative approach: Fix typo or use keyword, ‘neuopathy’ to include case.
- Note ID 18 is false negative because neuropathy has been described as ‘diabetic nerve pain’. Alternative approach: Use ‘diabetic nerve pain’ as keyword to include case.
- Note ID 85. Cause of false negative unknown.
- Note ID 108. Cause of false negative unknown.
- Note ID 135. ‘Retinopathy’ has been described as ‘optic nerve damage’. Alternative approach: Use keyword ‘optic nerve damage’ to include case.