Statistical Inference for NLP algorithms

Kishore Basu & TJ Ayoub

What are the authors trying to solve?

Computer Science/Linguistics

Statistical Approaches

- In NLP, statistical techniques used infrequently
- Uncertainty quantification is difficult
- word2vec has strong statistical underpinnings

"History Diabetes high blood pressure"

$C = c$ (Center Word)	$C' = c'$ (Neighbouring Word)	$D = 0/1$ (Boolean neighbour)
diabetes	history	$D = 1$
diabetes	diabetes	$D = 1$
diabetes	high	$D = 1$
diabetes	blood	$D = 1$
diabetes	pressure	$D = 0$
↓	↓	↓
pressure	history	$D = 0$

$PMI > 0$

$- C, C'$ *associated*

$PMI < 0$

$-C, C'$ *not associated*

$$PMI = log \frac{P(C = c \mid C' = c', D = 1)}{P(C = c \mid D = 1)}$$

?

Can statistical techniques improve our understanding of diabetes classification?

Yes, but we need to do some work first.

- PMI model
- MLE estimation
- Multivariate delta method

$$PMI = log \frac{P_1(C = c \,|C' = c', D = 1)}{P_2(C = c \,|D = 1)}$$

$$\begin{array}{lcccccc} & & \text{MLE} & & \text{MVDM} & & \text{DM} \\ P_1(C = c \,|C' = c', D = 1) = \dfrac{\exp(\beta^T X)}{1 + \exp(\beta^T X)} & \rightarrow & \hat{\beta} & \rightarrow & \widehat{\sigma_1}^2 = Var(\widehat{P_1}) & \searrow & \\ & & & & & & \widehat{PMI} \\ P_2(C = c \,|D = 1) = \dfrac{\exp(\alpha)}{1 + \exp(\alpha)} & \rightarrow & \hat{\alpha} & \rightarrow & \widehat{\sigma_2}^2 = Var(\widehat{P_2}) & \nearrow & \end{array}$$

Building a predictive model from data

- Goal: Identify type-2 diabetes in EHR
- Two summary statistics for two groups
- Word association pairs, words with the ‘diabet’ stem

$(diabet-, c')$

$m_k PMI_d^d$

$m_k PMI_d^{xd}$

(c, c')

$m_k PMI^d$

$m_k PMI^{xd}$

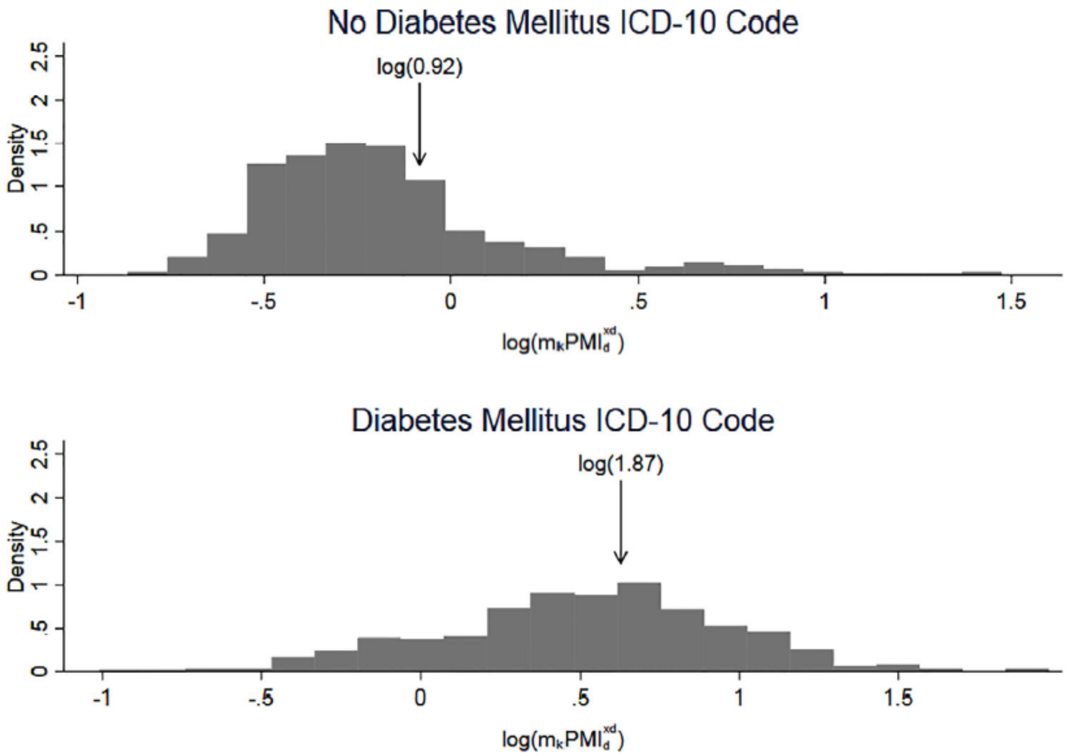
RESULTS

1000 patients
500 Diabetes, 500 non-diabetes

61,489 total words
Each patient has many records (notes)

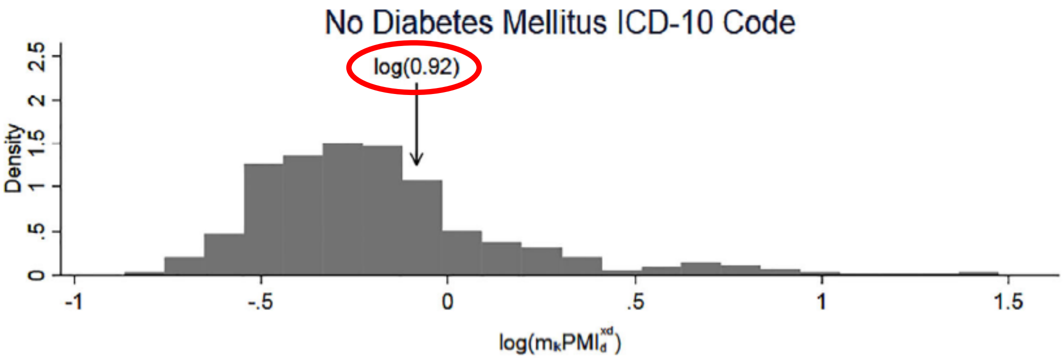
1500 words kept
Reduces dimensionality of dataset

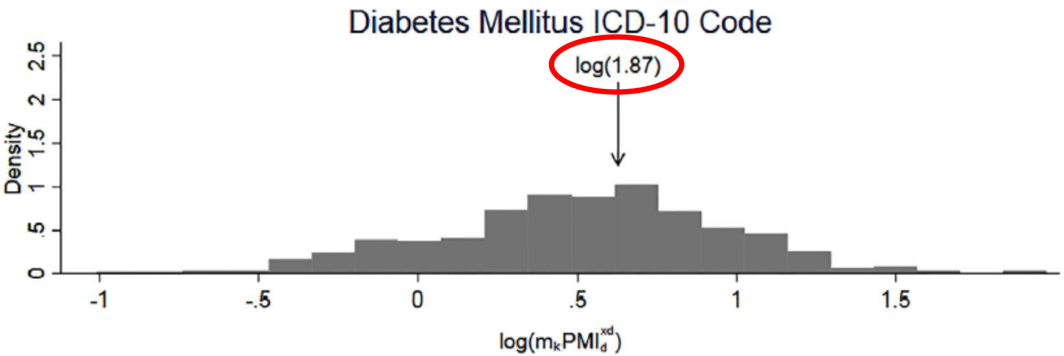
PREDICTIVE MODEL



Mean PMI over non-diabetics where $C' = `diabet'$

PREDICTIVE MODEL





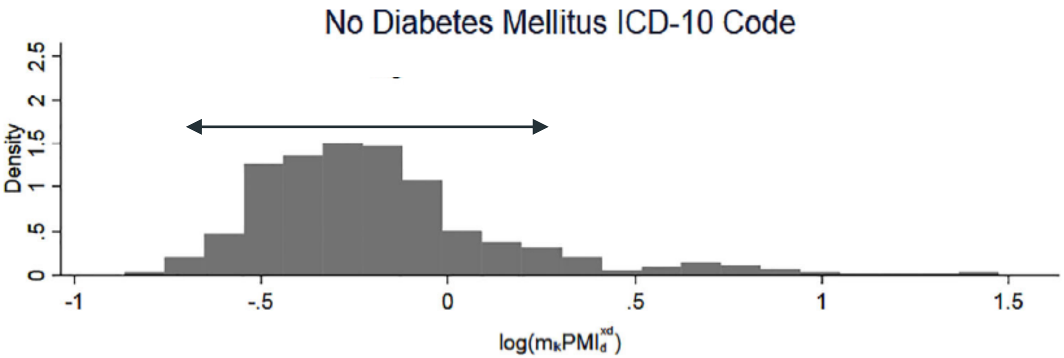
Mean PMI over non-diabetics where $C' = \text{`}\boldsymbol{diabet}'$

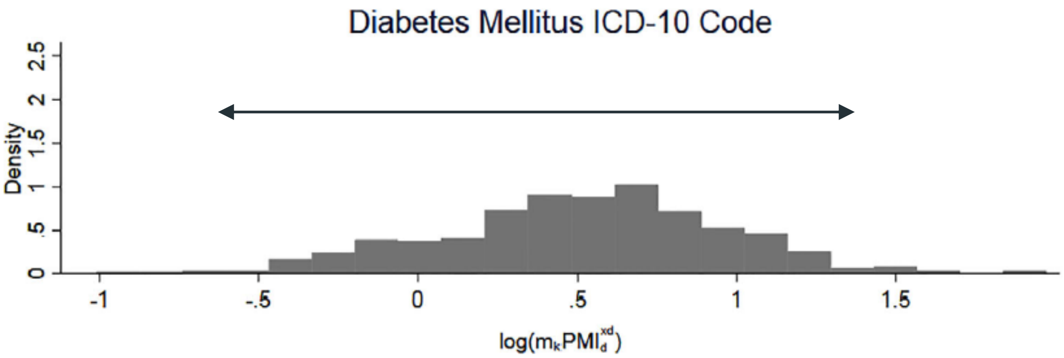
- **Key Takeaway:**

Less words associated with ‘diabet’ in non-diabetics

More words associated with ‘diabet’ in diabetics

PREDICTIVE MODEL





Mean PMI over non-diabetics where $C' = `diabetes'$

- **Key Takeaway:**

Non-diabetics have tighter distribution

PMI's in non-diabetes group are better at predicting who doesn't have diabetes

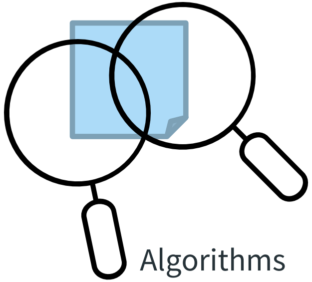
So understanding the distribution of the PMI can tell us a lot!

In summary,

- Novel framework to calculate SEs for PMI

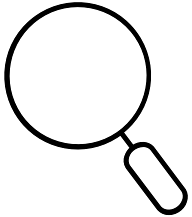
$$PMI = log \frac{P(C = c \mid C' = c', D = 1)}{P(C = c \mid D = 1)} \rightarrow \frac{\widehat{PMI}}{Var(\widehat{PMI})}$$

- Importance of statistical analysis in NLP and data science



Computer Science

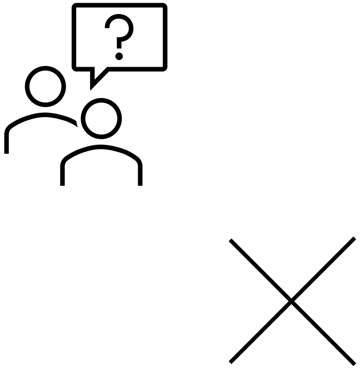
Algorithms

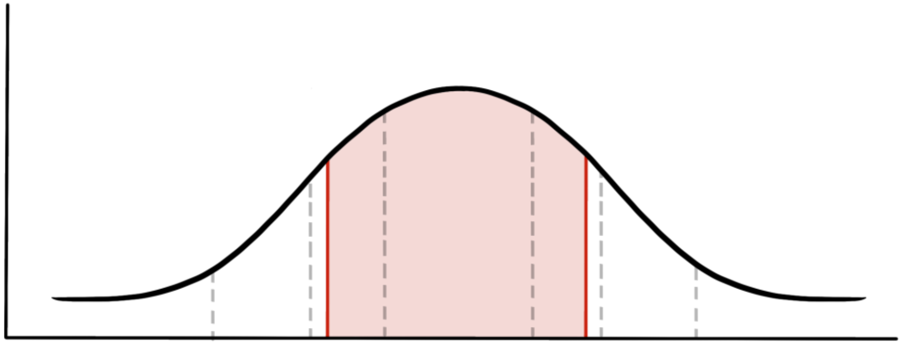


Statistical theory

In summary,

- High relevance contribution in a sparse literature space





Thank you for listening