AUTOMATICALLY GRADING BRAZILIAN STUDENT ESSAYS

Erick Fonseca, Ivo Medeiros, Dayse Kamikawachi, Alessandro Bokan

Letrus

PROPOR 2018

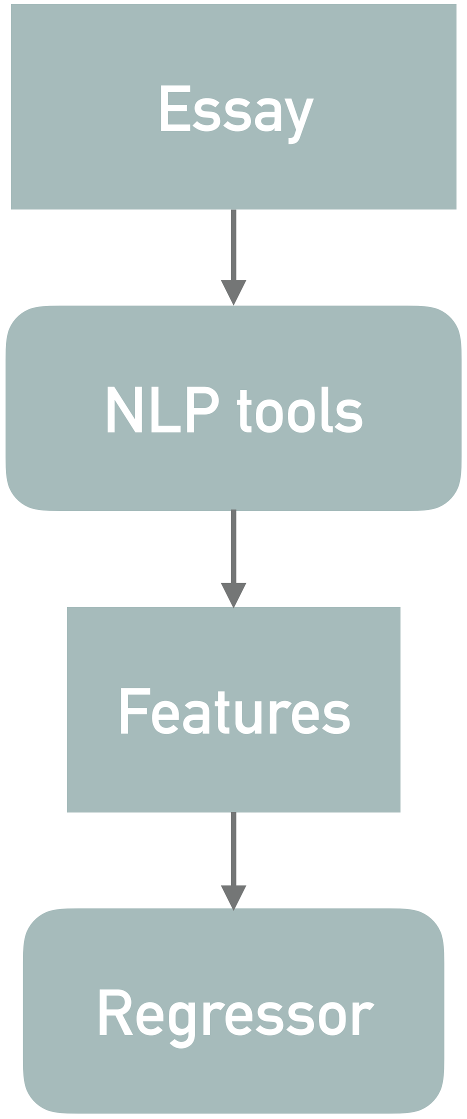
Letrus

AUTOMATIC ESSAY SCORING (AES)

- Score students essays — somewhat subjective!
- Fast, cheap and **deterministic**
- Can be exploited by students
  - Good for feedback during writing practice

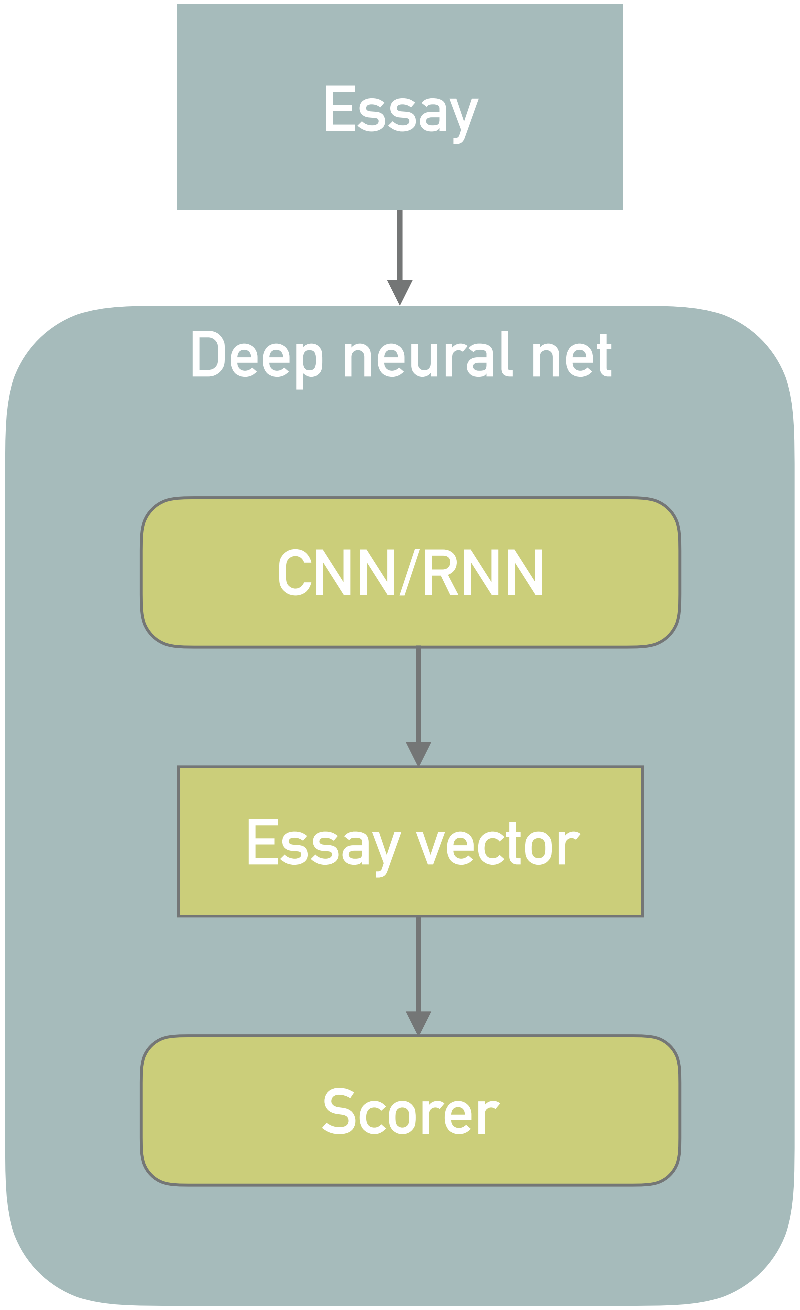
AES APPROACHES

- Early AES systems trained regressors with a large number of features:
  - Counts of words
  - POS tags
  - Syntactic structures
  - Named entities
  - n-grams
  - Spelling and grammar mistakes
  - etc…



AES APPROACHES

- More recently, neural networks
  - Create a vector representation for the essay
  - Learn a scorer
- Different architectures:
  - CNNs or RNNs
  - Single level or sentence level followed by text level



LET'S TRY BOTH!

- We tried both **neural networks** and **feature-based** models
  - Compare their pros and cons!

- We used a dataset of ~56k essays graded by humans
  - Larger than the English benchmark!

AES IN PORTUGUESE — ENEM

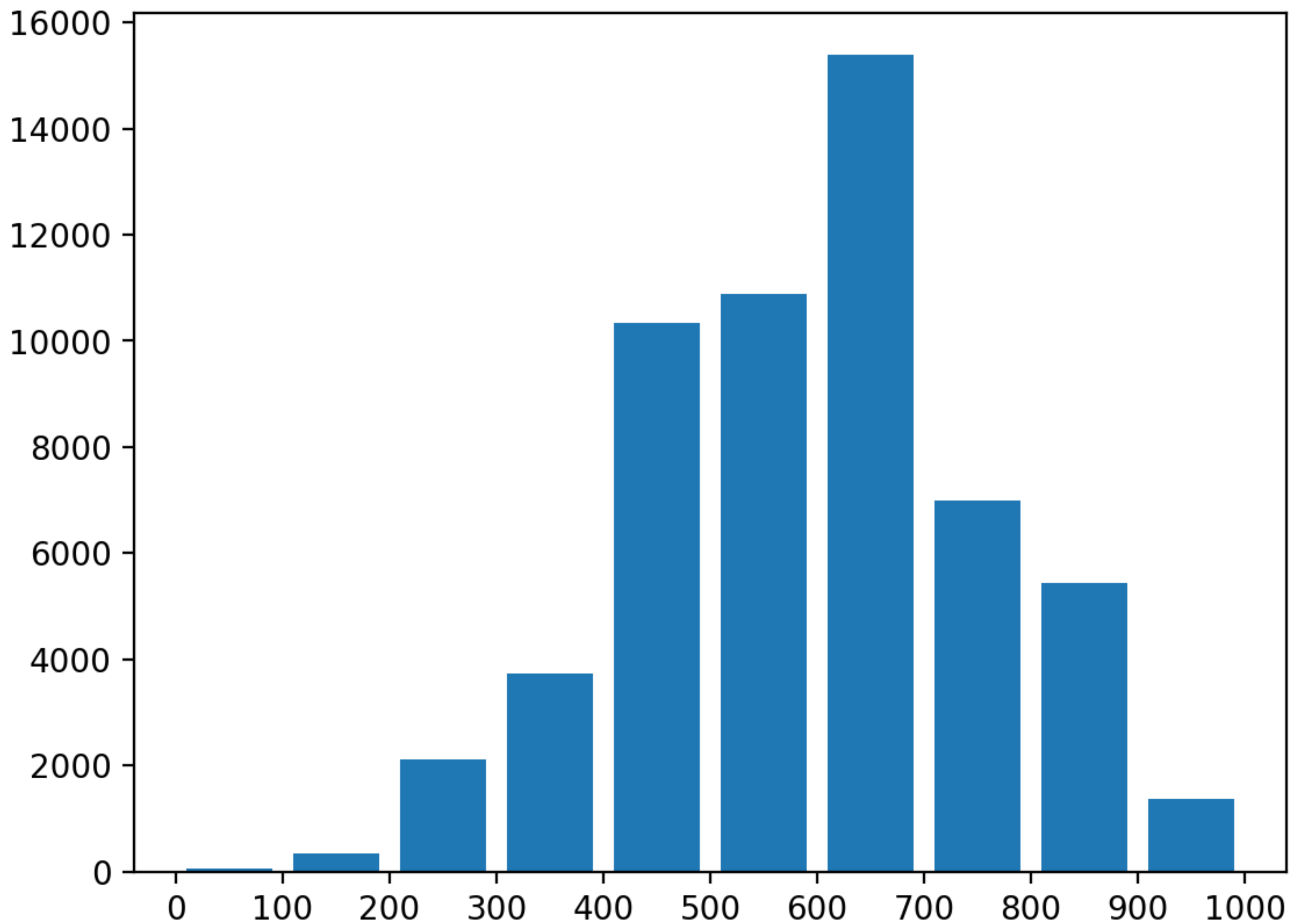
- Exam for high school students
  - Argumentative essays with a given topic

- ENEM scores essays in five competencies:

1. Standard written norm
2. Adherence to the topic and style
3. Defend a point of view
4. Usage of argumentative language
5. Proposal of a solution for the given problem

ENEM DATASET

- Each competency is scored from 0 to 200
  - Total essay score from 0 to 1000
- Scores have a gaussian distribution



HOW OUR DATA LOOKS LIKE

Metric	Mean (sd)
Tokens / sentence	32.0 (±18.2)
Tokens / essay	329.2 (±101.4)
Sentences / paragraph	2.4 (±1.3)
Sentences / essay	10.3 (±4.3)
Paragraphs / essay	4.3 (±1.0)

DEEP NEURAL NETWORK

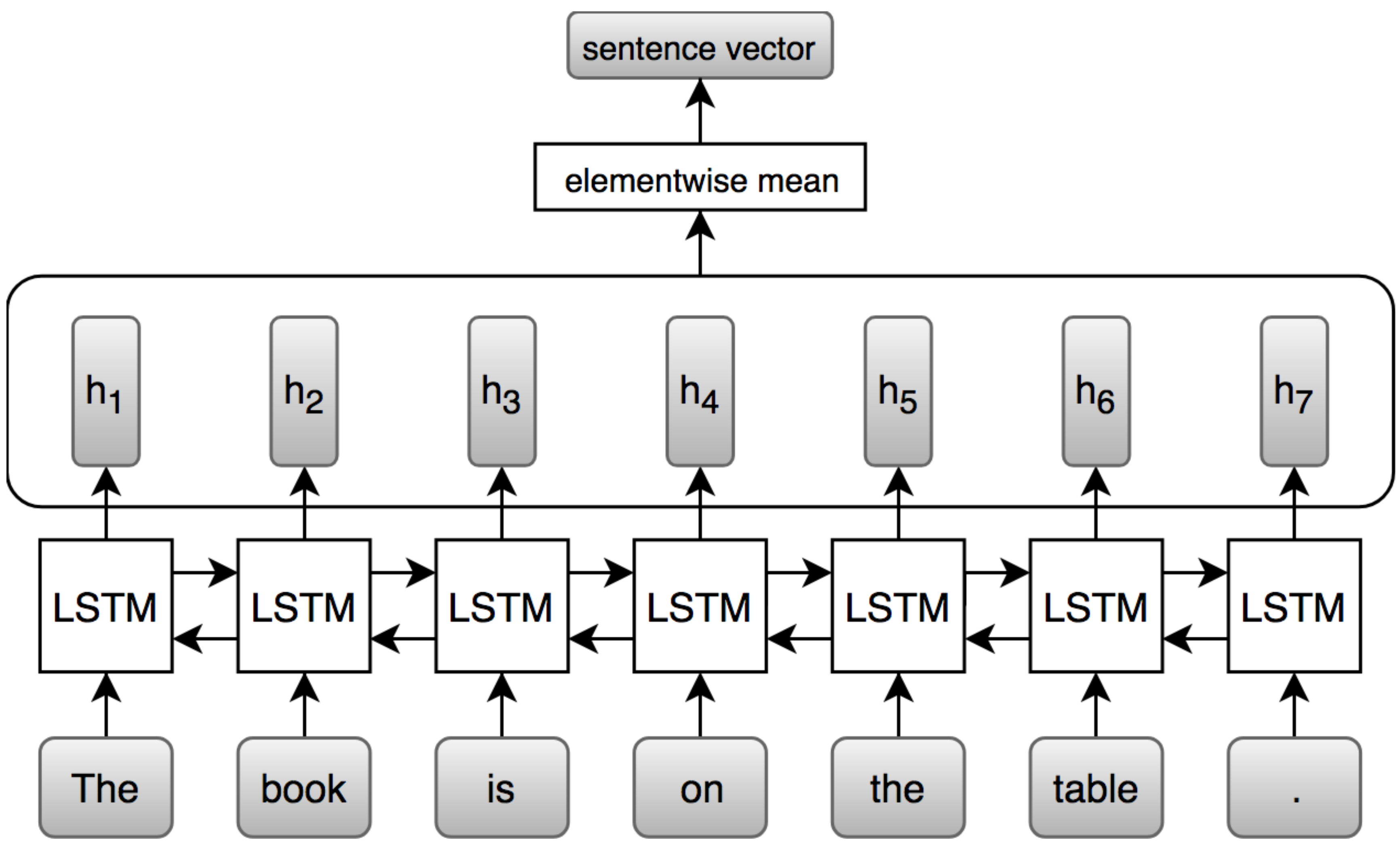
- The good:
  - Simpler to design
    - No need to handcraft features
  - Can learn some subtleties which are hard to describe

- The bad:
  - Harder to train
    - Needs much more computational power
    - Careful parameter tuning

DEEP NEURAL NETWORK

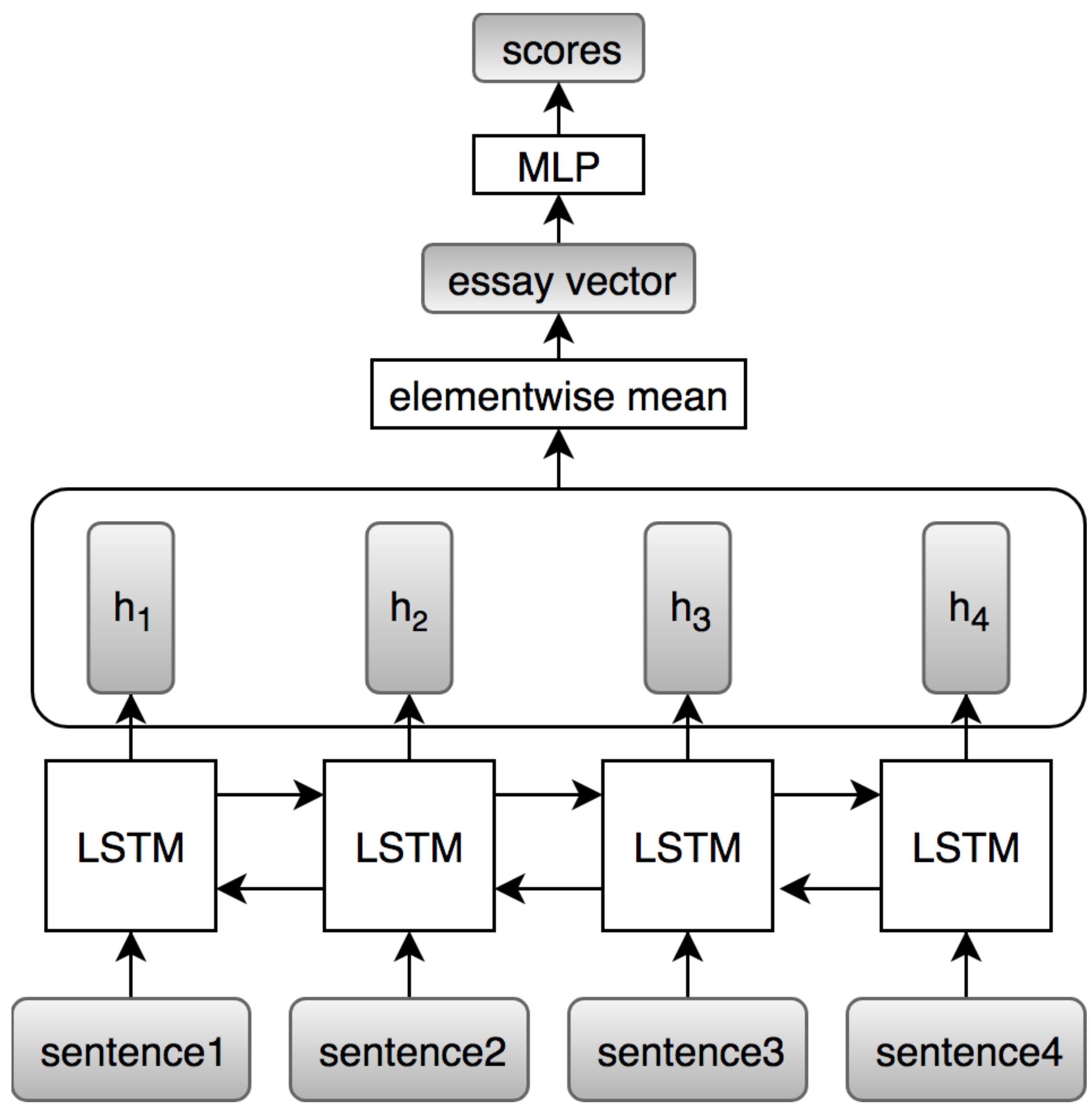
➤ Two levels of LSTMs

1. Read words and generate sentence vectors
2. Read sentences and generate an essay vector



DEEP NEURAL NETWORK

- Two levels of LSTMs
  1. Read word embeddings and generate sentence vectors
  2. Read sentences and generate an essay vector



DEEP NEURAL NETWORK

- Some variations yielded worse results
  - Max pooling instead of mean
  - CNNs instead of LSTMs

- The network outputs 5 scores
  - Sigmoid activation; normalize scores to range [0, 1]
  - Extra hidden layers did not help
  - Optimize the Mean Squared Errors: $\sum_{i}^{5} (y_i - \hat{y}_i)^2$

FEATURE ENGINEERING

- The bad:
  - Hard to design
    - Try to explain what makes an essay great!
  - Needs more preprocessing tools

- The good:
  - Computationally faster
  - Easier to interpret

FEATURE ENGINEERING

- Only run a POS tagger
  - Parsing is challenging because of mistakes (future work!)
- Use a list of hand picked expressions
  - Connectives, propositives, oralities
- Use a list of automatically extracted words and n-grams
  - Appearing in 5-50% of the essays
  - Pearson $\rho \geq 0.1$ with scores

FEATURE ENGINEERING — FEATURES

- Extract a vector of 681 features:
  - Number of commas, characters, tokens, types, sentences, token/sentence ratio, OOV words, OOV types, words from the prompt (…)
  - Presence of words and phrases from the handcrafted lists
  - Presence of relevant words and n-grams
  - Counts and ratios of each POS tag
  - Presence of relevant POS tag n-grams

- For each competency, only keep features with $\rho \geq 0.1$

EXPERIMENTAL SETUP

- Two metrics:
  - **Quadratic Weighted Kappa (QWK)** — Popular metric for AES; but disregards the error magnitude
  - **Root Mean Squared Error (RMSE)** — More appropriate for regression

- We compare with Amorim & Veloso (2017)
  - Only other work in Portuguese
  - … but with another and smaller corpus

Letrus

RESULTS

Model	C1	C2	C3	C4	C5	Total
Gradient Boosting	**25.81**	**26.02**	**27.40**	**28.34**	41.19	**100.00**
Linear Regression	26.10	26.37	27.75	28.42	42.07	101.53
Deep Network	27.75	26.58	27.51	29.26	**38.85**	100.59
Average baseline	38.26	33.53	34.72	39.47	55.27	160.42

RMSE (lower is better)

Model	C1	C2	C3	C4	C5	Total
Gradient Boosting	**0.676**	**0.511**	**0.508**	**0.619**	0.577	**0.752**
Linear Regression	0.667	0.499	0.493	0.615	0.564	0.747
Deep Network	0.615	0.503	0.500	0.508	**0.636**	0.750
Average Baseline	0	0	0	0	0	0
Amorim & Veloso	0.315	0.268	0.231	0.270	0.139	0.367

QWK (higher is better)

CONCLUSIONS

- Feature engineering models are better at C1-4
  - Easier competencies to describe how to score

- Competency 5 is the most difficult to score
  - Neural networks are better at it
    - … because of subjectivity?
  - Our models are more stable across competencies than Amorim & Veloso
- RMSE makes clear which competencies are harder

CONCLUSIONS

- AES is still incipient in Portuguese!
- Feature-based models and DNNs have comparable performance
- Many interesting directions for future works!
  - Parsing (for grammatically incorrect sentences)
  - Other network architectures
  - Evaluate students’ writing skill evolution

THANK YOU!
QUESTIONS?

erick@letrus.com.br

ivopdm@letrus.com.br

dayse@letrus.com.br

abokan@letrus.com.br