Universidade do Minho

APRENDIZAGEM PROFUNDA

David Teixeira PG55929
Eduardo Cunha PG55939
Nuno Rodrigues PG55966
Tiago Rodrigues PG56013

INTRODUÇÃO

- **Objetivo:** Distinguir textos gerados por IA de textos humanos, face à crescente sofisticação dos modelos de geração.
- Treino e comparação de diferentes modelos de Deep Learning.
- Avaliação do desempenho dos modelos.

CRIAÇÃO DO DATASET

- Recolha de dados de fontes públicas (artem9k/ai-text-detection-pile, NicolaiSivesind/human-vs-machine, ...)
- Pré-processamento, incluindo padronização das labels ("human" = 0, "AI" = 1), tokenização e remoção de duplicados
- Uso do CountVectorizer para converter textos em representações numéricas com 10.000 features
- Divisão do dataset em treino (70%), validação (10%) e teste (20%)
- Armazenamento dos dados processados em formato CSV

MODELOS RAIZ

Regressão Logística:

- Otimização com SciPy e Gradient Descent
- Regularização

DNNs:

- Funções de ativação (ReLU, Tanh, Sigmoid).
- Regularização (Dropout, L1/L2).
- Otimizadores (Adam, RMSProp).
- Diferentes Métricas (Accuracy e F1-score)
- Callbacks

RNNs:

- Problemas com dimensionalidade (1D vs. 3D).
- Embeddings adaptados (baixo desempenho).

MODELOS COM TENSORFLOW

DNNs:

- Melhoria com TF-IDF vs. tokenização simples (94% accuracy).
- Embeddings: GloVe (pré-treinado) vs. treino do zero.

RNNs:

- Comparação: SimpleRNN (dissipação) < GRU (eficiente) < LSTM (melhor desempenho).

Transformers:

- Alto custo computacional (1 hora e 20min/epoch).
- BERT: resultados modestos vs. complexidade.

ABORDAGENS COM LLMS

Zero/One-Shot Learning:

- Classificação por similaridade de embeddings (exemplo único).
- Bart, RoBERTa, DistilRoBERTa - Zero Shot
- Bert, RoBERTa, Bart - One Shot

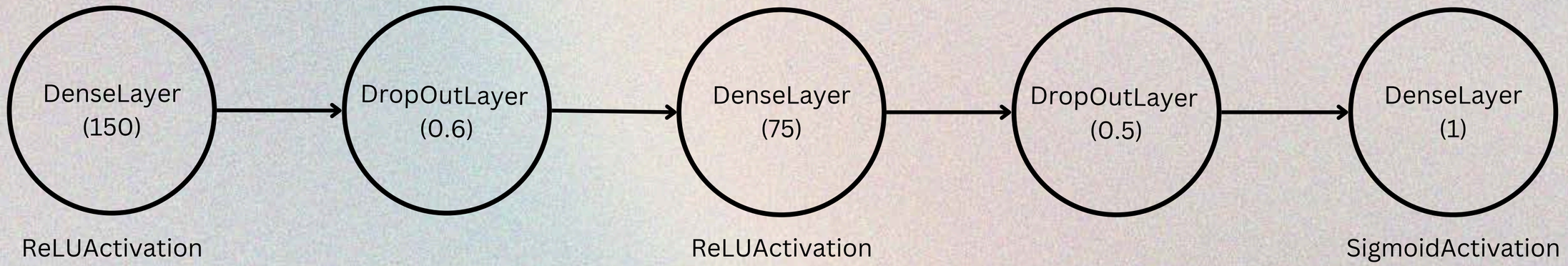
Engenharia de Prompts:

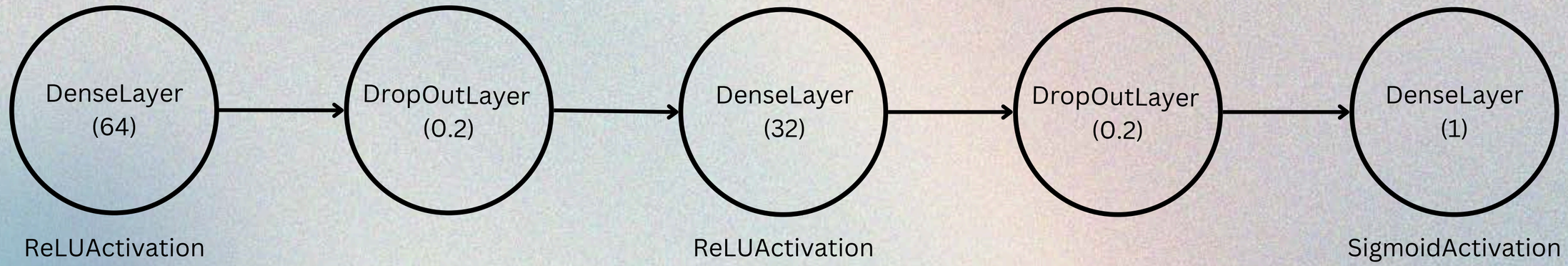
- Few-shot com BERT e GPT (instabilidade nos resultados).

RAG:

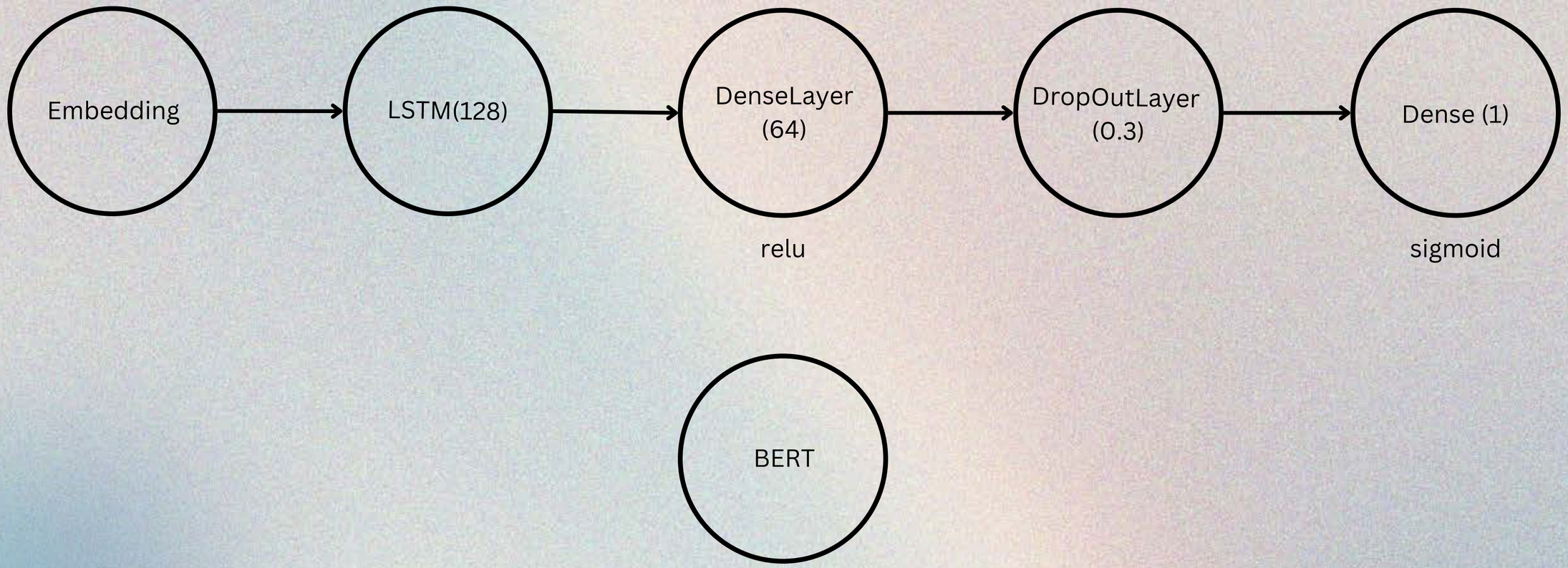
- Não implementado (complexidade desnecessária para o problema).

1 SUBMISSÃO


DenseLayer
(150)
DropOutLayer
(0.6)
DenseLayer
(75)
DropOutLayer
(0.5)
DenseLayer
(1)
ReLUActivation
ReLUActivation
SigmoidActivation


DenseLayer
(64)
DropOutLayer
(0.2)
DenseLayer
(32)
DropOutLayer
(0.2)
DenseLayer
(1)
ReLUActivation
ReLUActivation
SigmoidActivation

2 SUBMISSÃO


Embedding
LSTM(128)
DenseLayer
(64)
relu
DropOutLayer
(0.3)
Dense (1)
sigmoid
BERT

RESULTADOS

1ª Submissão (Modelos Raiz):

- 2º lugar (73% accuracy).

2ª Submissão (Modelos Tensorflow)

- 5º lugar (67 e 69% accuracy).

3ª Submissão (Modelos Tensorflow)

- TBD

CONCLUSÃO

Desenvolvemos modelos próprios e baseados em bibliotecas consolidadas para distinguir entre textos escritos por humanos e gerados por inteligência artificial. Explorámos diferentes abordagens, implementámos soluções e analisámos os resultados obtidos.

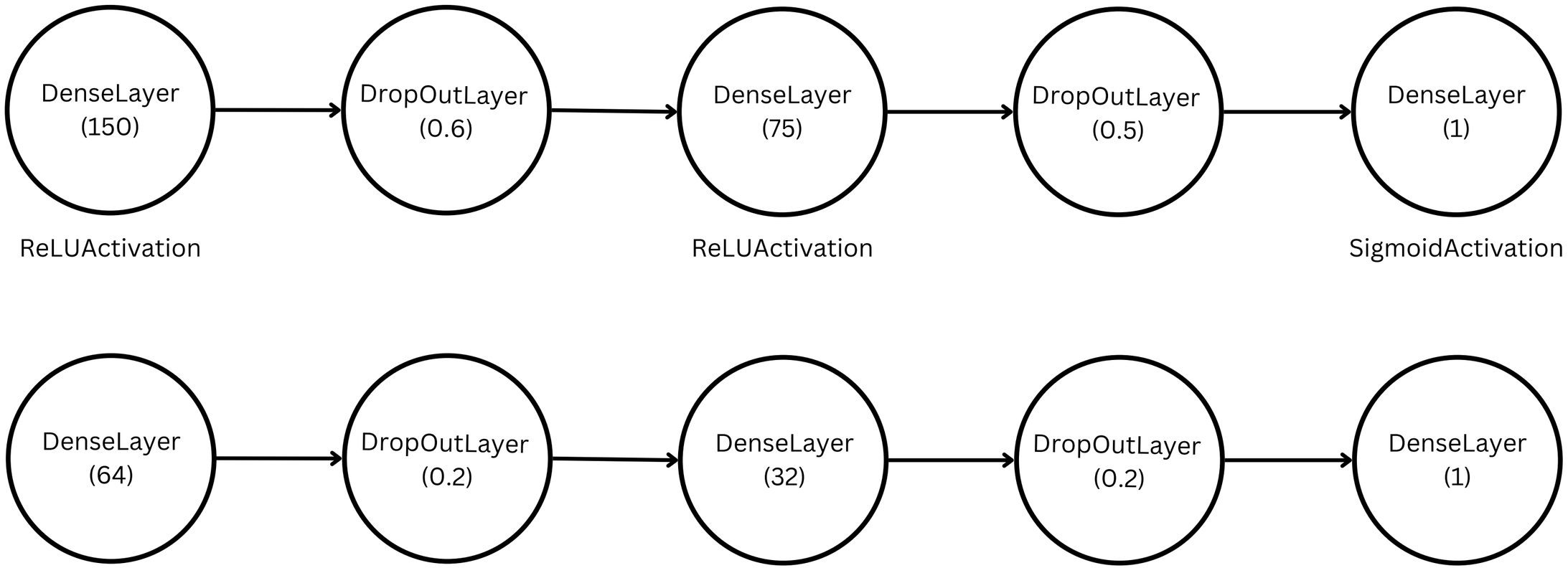
Cumprimos os objetivos propostos, criando modelos robustos e eficazes para um problema atual e em constante investigação. No contexto da competição, os resultados foram bastante competitivos, refletindo a qualidade do estudo e da preparação.

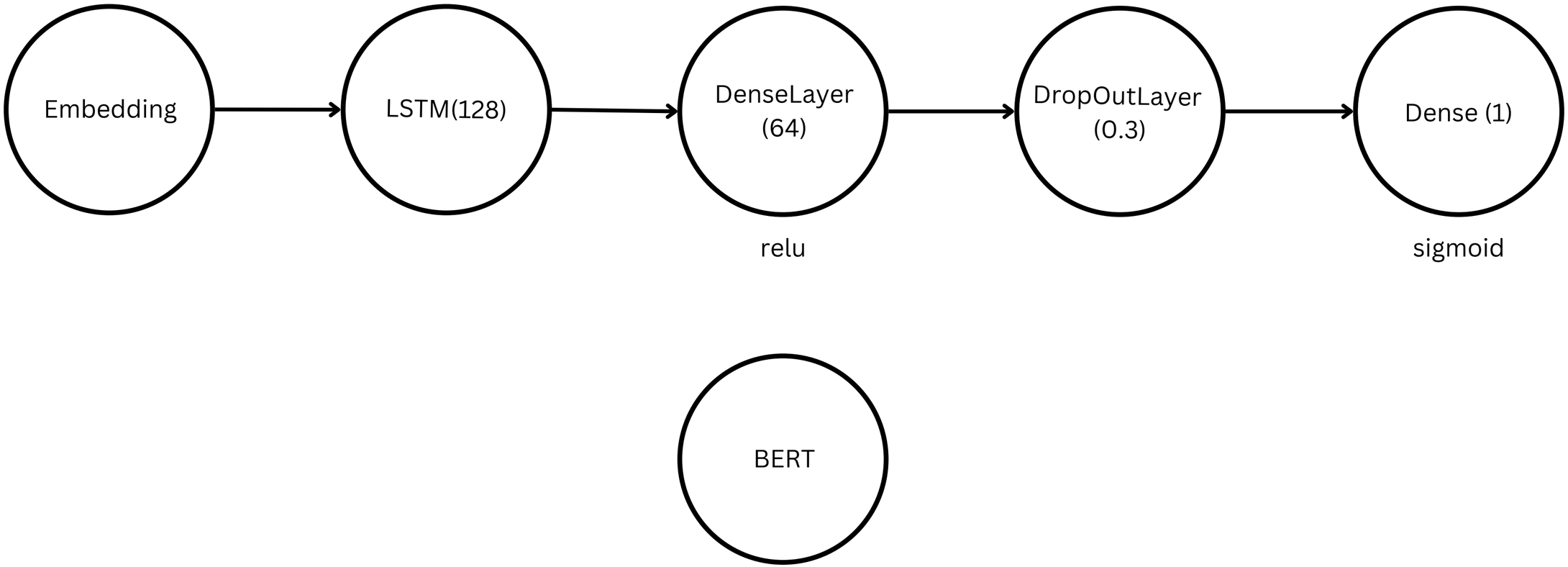
Este trabalho também permitiu consolidar e aprofundar os conceitos estudados, aplicando-os na prática e explorando novas perspetivas sobre o tema.

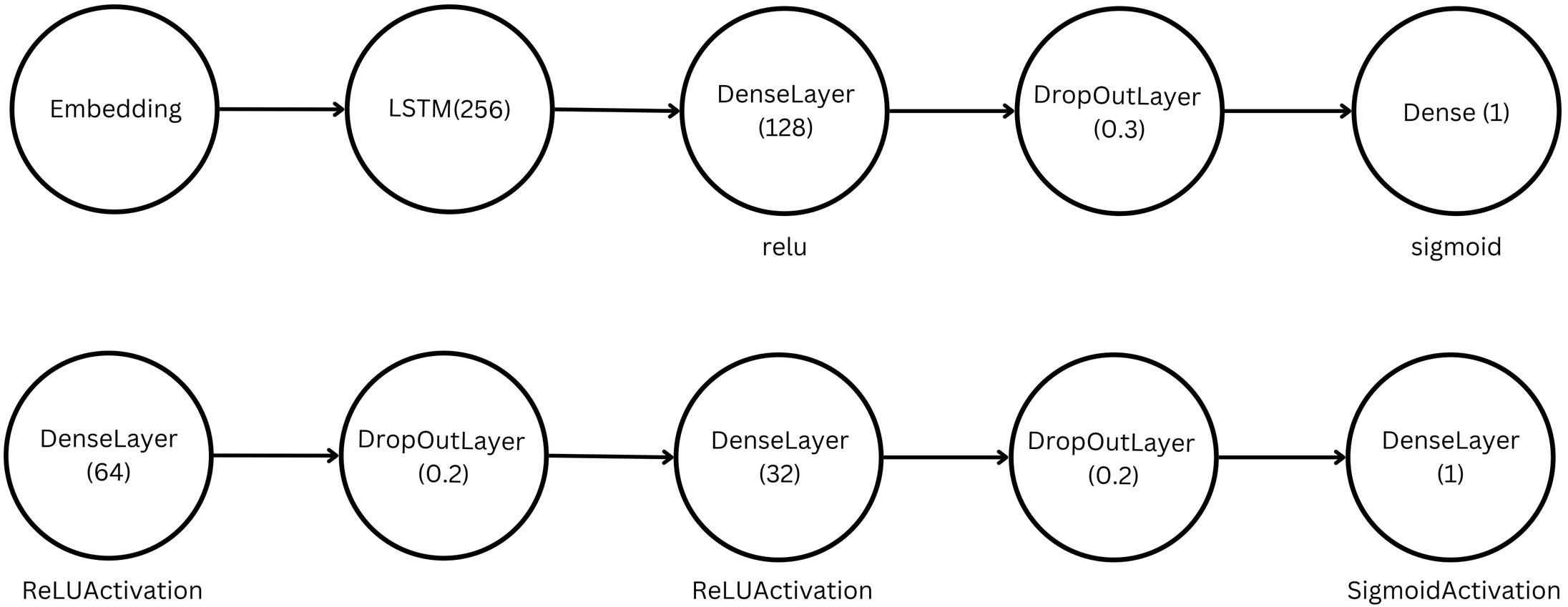
Universidade do Minho

APRENDIZAGEM PROFUNDA

David Teixeira PG55929
Eduardo Cunha PG55939
Nuno Rodrigues PG55966
Tiago Rodrigues PG56013


DenseLayer
(150)
DropOutLayer
(0.6)
DenseLayer
(75)
DropOutLayer
(0.5)
DenseLayer
(1)
ReLUActivation
ReLUActivation
SigmoidActivation
DenseLayer
(64)
DropOutLayer
(0.2)
DenseLayer
(32)
DropOutLayer
(0.2)
DenseLayer
(1)


Embedding
LSTM(128)
DenseLayer
(64)
relu
DropOutLayer
(0.3)
Dense (1)
sigmoid
BERT


Embedding
LSTM(256)
DenseLayer
(128)
relu
DropOutLayer
(0.3)
Dense (1)
sigmoid
DenseLayer
(64)
ReLUActivation
DropOutLayer
(0.2)
DenseLayer
(32)
ReLUActivation
DropOutLayer
(0.2)
DenseLayer
(1)
SigmoidActivation