Universidade de Brasília - UnB
Departamento de Ciência da Computação
Programa de pós-graduação em Computação Aplicada (PPCA)
Mineração de Dados Massivo (MDM)

AVALIAÇÃO DO DESEMPENHO DE ALGORITMOS DE ML E IA ASSOCIADOS A FEATURES DE TOKEN E FUZZY

Augusto Samuel Modesto
Marcus Fabricio Ferreira Paula

Agenda

- **Introdução**
  - Contexto
  - Objetivo
  - Hipóteses
- **Trabalhos Relacionados**
- **Conjunto de Dados Quora de Questões Similares**
- **Visão Geral e Abordagem**
  - Abordagem
  - Reamostragem
  - Pré-Processamento
  - Engenharia de Features e Vetorização
  - Modelos
  - Métricas de Avaliação dos Modelos
- **Resultados e Discussões**
- **Conclusões e Trabalhos Futuros**

Introdução

Contexto

- A internet possibilitou interação entre as pessoas de todo o mundo
- Surgimento das plataforma de perguntas e respostas (CQAs)
- Participantes fazem perguntas e podem responder de forma prática
- Tornou-se fácil encontrar respostas para as perguntas

Contexto

- A praticidade de gerar perguntas nestas plataformas, gera um grande volume de perguntas repetidas

- As perguntas repetidas, geram esforço da própria comunidade de usuários, em respondê-las.

Objetivo

Classificar perguntas como repetida, de forma automática

Hipóteses

(1) É possível ter resultados melhores incluindo à vetorização, recursos (features) de token e fuzzy dos pares similares

(2) Redes neurais tem performance melhor que algoritmos lineares e árvores de decisão neste contexto

Trabalhos Relacionados

Trabalho relacionados

- Identificação algumas características lexicais e sintáticos das perguntas
- Utilização da rede Continuous Bag of Words
- Ranqueamento e classificação de perguntas
- Utilização de redes LTSM
- Abordagem ontológica (Web Ontology Language - OWL)
- Contexto da medicina - Covid-19

Dataset Quora de perguntas similares

Dataset Quora

- Dados da Plataforma Quora
- Kaggle Quora Question Pairs
- Conjunto de treino com **404.290 pares de perguntas**
- Conjunto de teste com **2.345.795 pares de perguntas**
- Perguntas estão em inglês

Dataset Quora

- Campos que compõem o banco de dados:

1. Id
2. Qid1
3. Qid2
4. Question1
5. Question2
6. Is_duplicate

Dataset Quora

- Campos que compõem o banco de dados:

1. Id
2. Qid1
3. Qid2
4. Question1
5. Question2
6. Is_duplicate

id	qid1	qid2	question1	question2	isduplicate
0	1	2	What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
5	11	12	Astrology: I am a Capricor Sun Cap moon and cap rising... what does that say about me?	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?	1

Visão Geral da Abordagem

Abordagem

A abordagem passa pelas seguintes fases:

a) Reamostragem do *dataset*

b) Pré-Processamento

c) Engenharia de Features e Vetorização

d) Modelos

e) Métricas de Avaliação dos Modelos

Reamostragem do Dataset

- Dados originais desbalanceados:
  - 63% Não Similares
  - 37% Similares

- Etapa 1: Reamostragem para equilibrar as classes
  - Extraído 149.263 de cada conjunto de classe
  - Os dados foram embaralhados
  - Novo conjunto com 298.526, sendo:
    - 50% Não Similares
    - 50% Similares

- Etapa 2: Separação do dataset em *train* e *test*

Pré-Processamento

- Redução do vocabulário
- Tratamento de dados faltantes
- Conversão do texto das questões para minúsculos
- Remoção de caracteres inválidos e *stop word*
- Tratamento de caracteres especiais
- Descontração de palavras, comum no vocabulário inglês

Pré-Processamento

Antes:

id	qid1	qid2	question1	question2	isduplicate
0	1	2	What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
5	11	12	Astrology: I am a Capricor Sun Cap moon and cap rising... what does that say about me?	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?	1

Depois:

	Pergunta 1	Pergunta 2	Duplicado
0	step step guide invest share market india	step step guide invest share market	0
1	astrology capricor sun cap moon cap rising say	triple capricorn sun moon ascendant capricorn say	1

Engenharia de Features e Vetorização

- Criação de conjuntos de categorias de tokens, sendo:
  - Contadoras
  - Proporções
  - Condicionais
  - Comprimento

- Criação de Conjunto features Fuzzy

Engenharia de Features e Vetorização

- Criação de conjuntos de categorias de tokens, sendo:
  - **Contadoras**
    - Quantidade de tokens
    - Tamanho das questões
    - Tokens em comum
    - Adjetivos em comum
    - Nomes próprios em comum
    - Substantivos em comum
  - Proporções
  - Condicionais
  - Comprimento

Engenharia de Features e Vetorização

- Criação de conjuntos de categorias de tokens, sendo:
  - Contadoras
  - **Proporções**
    - Proporção número de tokens sobre comprimento da menor pergunta
    - Proporção número de tokens sobre comprimento da maior pergunta
  - Condicionais
  - Comprimento

Engenharia de Features e Vetorização

- Criação de conjuntos de categorias de tokens, sendo:
  - Contadoras
  - Proporções
  - **Condicionais**
    - Primeiros tokens iguais
    - Últimos tokens iguais
  - Comprimento

Engenharia de Features e Vetorização

- Criação de conjuntos de categorias de tokens, sendo:
  - Contadoras
  - Proporções
  - Condicionais
  - **Comprimento**
    - Média do comprimento das perguntas
    - Mediana do comprimento das perguntas
    - Razão entre maior e menor perguntas
    - Diferença absoluta entre as perguntas

Engenharia de Features e Vetorização

- Criação de conjuntos de categorias de tokens, sendo:
  - Contadoras
  - Proporções
  - Condicionais
  - Comprimento

- **Criação de Conjunto features Fuzzy**
  - Parcial Fuzzy
  - Token sort ratio
  - Token set ratio

Engenharia de Features e Vetorização

- Vetorização: captura de significados semânticos de frases
  - TF-IDF
  - Média Ponderada IDF do Word2Vec

Engenharia de Features e Vetorização

- Vetorização: captura de significados semânticos de frases
  - **TF-IDF**
    - 59.457 dimensões por pergunta
    - Totalizando 118.914 dimensões
  - Média Ponderada IDF do Word2Vec

Engenharia de Features e Vetorização

- Vetorização: captura de significados semânticos de frases
  - TF-IDF
  - **Média Ponderada IDF do Word2Vec**
    - 300 dimensões por pergunta
    - Totalizando 600 dimensões

Modelos

- Algoritmos
  - Regressão Logıística
  - XGBoost
  - Rede Siamesa
- Modelos de dados são uma combinação
  - Features
  - Vetores
  - Algoritmos

Modelos

Nº Modelo	Características	Algoritmo	Dimensões
Modelo 1	Features (token e fuzzy)	LR e XGBoost	21
Modelo 2	TF-IDF	LR e XGBoost	118.914
Modelo 3	Média ponderada IDF do Word2Vec	LR e XGBoost	600
Modelo 4	TF-IDF + Features	LR e XGBoost	118.935
Modelo 5	Média ponderada IDF do Word2Vec + Features	LR e XGBoost	621
Modelo 6	TF-IDF	Rede Siamesa	118.914

Modelos

Nº Modelo	Características	Algoritmo	Dimensões
Modelo 1	Features (token e fuzzy)	LR e XGBoost	21
Modelo 2	TF-IDF	LR e XGBoost	118.914
Modelo 3	Média ponderada IDF do Word2Vec	LR e XGBoost	600
Modelo 4	TF-IDF + Features	LR e XGBoost	118.935
Modelo 5	Média ponderada IDF do Word2Vec + Features	LR e XGBoost	621
Modelo 6	TF-IDF	Rede Siamesa	118.914

Modelos

Nº Modelo	Características	Algoritmo	Dimensões
Modelo 1	Features (token e fuzzy)	LR e XGBoost	21
Modelo 2	TF-IDF	LR e XGBoost	118.914
Modelo 3	Média ponderada IDF do Word2Vec	LR e XGBoost	600
Modelo 4	TF-IDF + Features	LR e XGBoost	118.935
Modelo 5	Média ponderada IDF do Word2Vec + Features	LR e XGBoost	621
Modelo 6	TF-IDF	Rede Siamesa	118.914

Modelos

Nº Modelo	Características	Algoritmo	Dimensões
Modelo 1	Features (token e fuzzy)	LR e XGBoost	21
Modelo 2	TF-IDF	LR e XGBoost	118.914
Modelo 3	Média ponderada IDF do Word2Vec	LR e XGBoost	600
Modelo 4	TF-IDF + Features	LR e XGBoost	118.935
Modelo 5	Média ponderada IDF do Word2Vec + Features	LR e XGBoost	621
Modelo 6	TF-IDF	Rede Siamesa	118.914

Modelos

Nº Modelo	Características	Algoritmo	Dimensões
Modelo 1	Features (token e fuzzy)	LR e XGBoost	21
Modelo 2	TF-IDF	LR e XGBoost	118.914
Modelo 3	Média ponderada IDF do Word2Vec	LR e XGBoost	600
Modelo 4	TF-IDF + Features	LR e XGBoost	118.935
Modelo 5	Média ponderada IDF do Word2Vec + Features	LR e XGBoost	621
Modelo 6	TF-IDF	Rede Siamesa	118.914

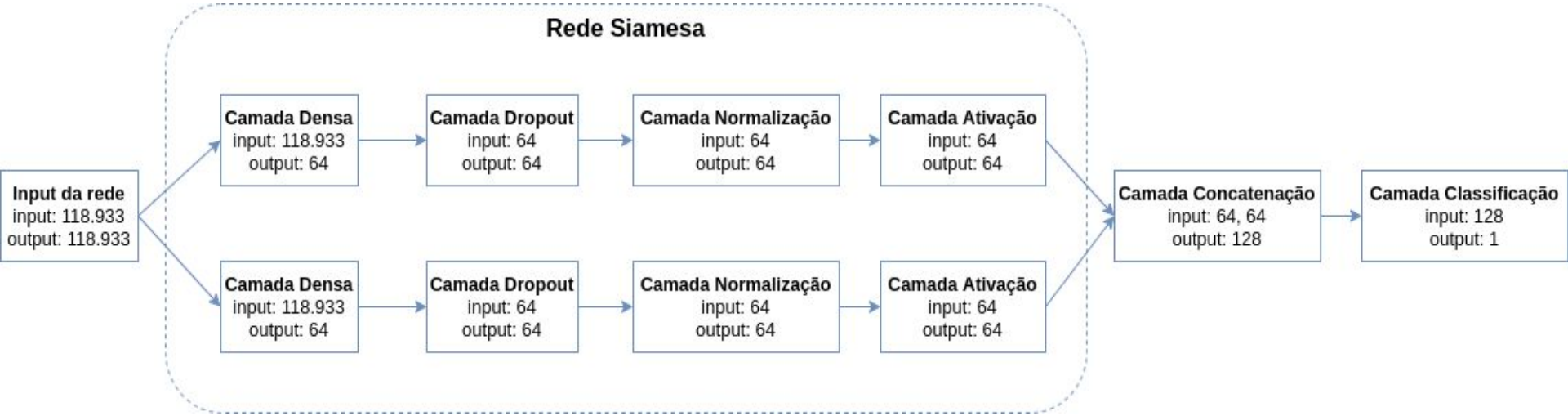
Modelos

Nº Modelo	Características	Algoritmo	Dimensões
Modelo 1	Features (token e fuzzy)	LR e XGBoost	21
Modelo 2	TF-IDF	LR e XGBoost	118.914
Modelo 3	Média ponderada IDF do Word2Vec	LR e XGBoost	600
Modelo 4	TF-IDF + Features	LR e XGBoost	118.935
Modelo 5	Média ponderada IDF do Word2Vec + Features	LR e XGBoost	621
Modelo 6	TF-IDF	Rede Siamesa	118.914

Modelos

- **Regressão Logística**
  - alfa = 0.00001
  - Regularização ElasticNet
- **XGBoost**
  - Estimadores: 100
  - Profundidade: 6
- **Rede Siamesa**
  - batchz size: 64
  - Epochs: 20
  - Optimizers: Adam
  - Learning_rate: 0.01

Modelos


Rede Siamesa
Input da rede
input: 118.933
output: 118.933
Camada Densa
input: 118.933
output: 64
Camada Dropout
input: 64
output: 64
Camada Normalização
input: 64
output: 64
Camada Ativação
input: 64
output: 64
Camada Densa
input: 118.933
output: 64
Camada Dropout
input: 64
output: 64
Camada Normalização
input: 64
output: 64
Camada Ativação
input: 64
output: 64
Camada Concatenação
input: 64, 64
output: 128
Camada Classificação
input: 128
output: 1

Métricas de Avaliação dos Modelos

- Acurácia
- F1-Score
- Log Loss

Resultados e Discussões

Resultados

	Regressão Logística			XGBoost		
	Log Loss	Acurácia	F1-Score	Log Loss	Acurácia	F1-Score
Modelo 1	0.55102	0.70486	0.71552	0.47443	0.76173	0.77820
Modelo 2	0.55030	0.71894	0.70863	0.58359	0.67565	0.60519
Modelo 3	0.65252	0.63148	0.63937	0.62977	0.68330	0.69850
Modelo 4	0.51556	0.74690	0.75843	0.49101	0.75968	0.78961
Modelo 5	**0.43078**	**0.79773**	**0.80273**	**0.44134**	**0.78143**	**0.79544**

	Rede Siamesa		
	Log Loss	Acurácia	F1-Score
Modelo 6	0.67036	0.75417	0.76461

Média ponderada IDF do Word2Vec + Features

Discussões

- Incremento de novas Features
  - Distância do cosseno
- Utilização do BERT e ROBERTa
- Melhorias na parametrização dos modelos
- Utilização de arquitetura de redes siamesas com LSTM

Conclusão e Trabalhos Futuros

Conclusões

Hipótese Aceita

(1) é possível ter melhores resultados incluindo à vetorização recursos (*feature*) de token e fuzzy dos pares de similaridade;

Hipótese Rejeitada

(2) redes neurais tem performance melhor que algoritmos lineares e árvores de decisão neste contexto.

Trabalhos Futuros

Avaliação dos modelos com técnicas de vetorização mais robustas como BERT e RoBerta

Adição de features pode melhorar o desempenho dos modelos.

Utilização de redes siamesas com LSTM, com parametrizações mais ajustadas

Obrigado!

Dúvidas?