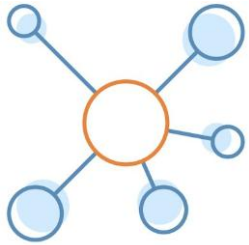
Aplicando Programação Genética na Geração de Classificadores de Sentimento

Airton Bordin Junior
airtonbjunior@gmail.com

Prof. Dr. Nádia Félix Felipe da Silva
nadia@inf.ufg.br

Prof. Dr. Celso Gonçalves Camilo Junior
celso@inf.ufg.br

Prof. Dr. Thierson Couto Rosa
thierson@inf.ufg.br



CBIC 2017

Roteiro

- Introdução
  - Contextualização
  - Problema
  - Objetivo
- Conceitos
- Materiais e Métodos
- Análise dos Resultados
- Conclusão

Introdução - Contextualização

- Web é comumente utilizada como plataforma para debates, opiniões, avaliações, entre outros
- Instituições, pessoas e empresas tem interesse em saber qual a opinião de um grupo de pessoas sobre determinado tema

Introdução - Contextualização

- A Análise de Sentimentos (AS) é uma linha de pesquisa que tem por objetivo a classificação das emoções de um determinado texto, geralmente como positivo, negativo ou neutro
- Duas abordagens principais: ***Machine Learning*** e **Análise Léxica**

Introdução - Contextualização

- Interesse do termo "*Sentiment Analysis*" – 2004 a 2017

Introdução - Problema

- [????] Aspectos inerentes ao contexto das opiniões que serão avaliadas
- Custo da construção de um classificador para um contexto específico (geralmente manual)

Introdução - Objetivo

- Formular a geração de um classificador de sentimento como um problema de busca e otimização
  - Encontrar um modelo dentro do espaço de modelos possíveis
- Geração de modelos eficientes de classificação de sentimentos usando Programação Genética

Roteiro

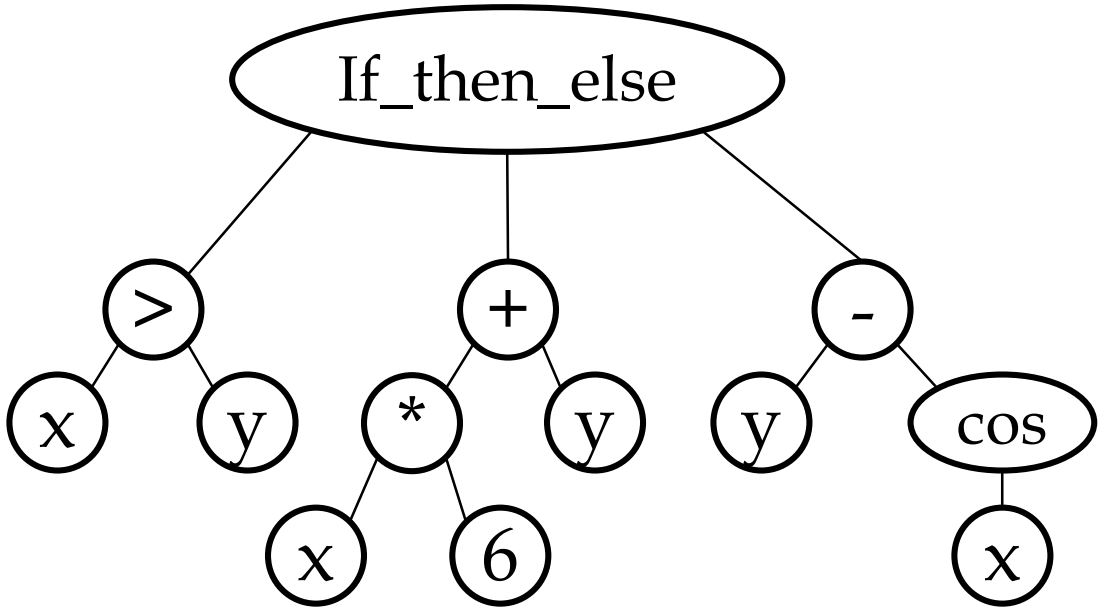
- Introdução
  - Contextualização
  - Problema
  - Objetivo
- Conceitos
- Materiais e Métodos
- Análise dos Resultados
- Conclusão

Programação Genética

- Resolução de problemas, de forma automatizada, sem demandar conhecimentos detalhados sobre a solução
- Programa: modelo capaz de, à partir de uma ou mais entradas, produzir uma saída para as mesmas

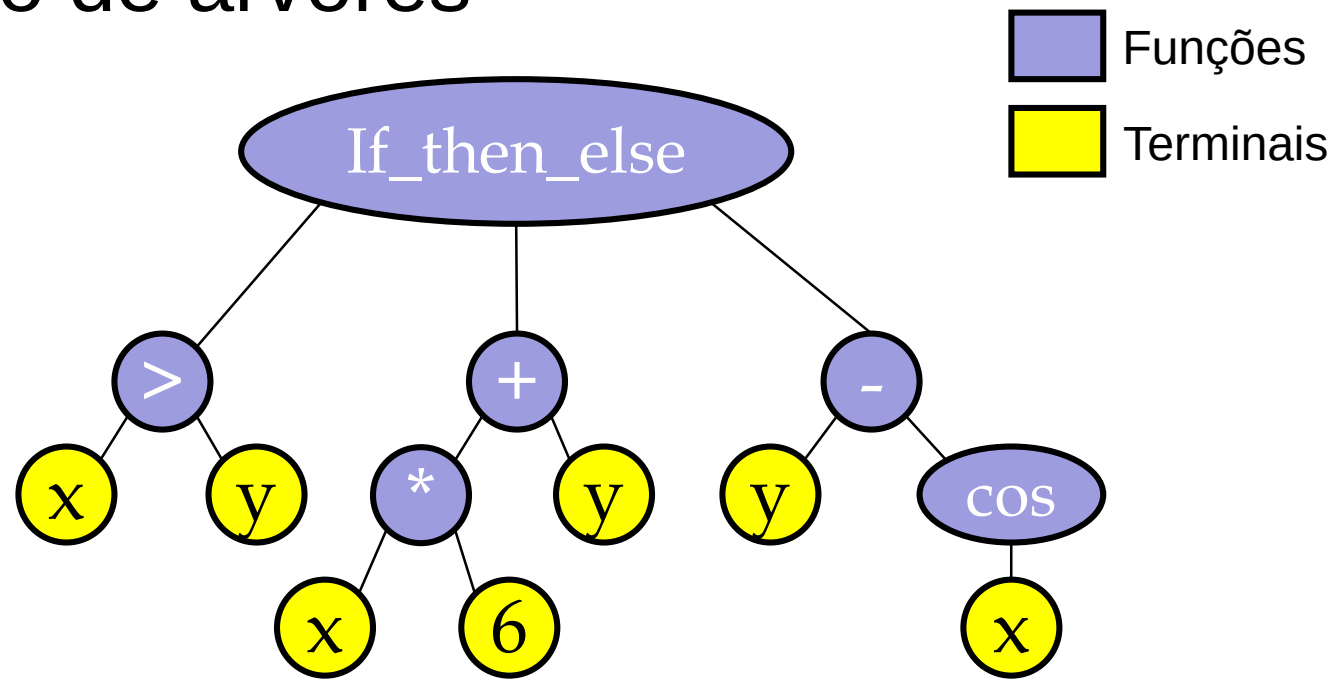
Programação Genética

- Modelos geralmente representados por meio de árvores

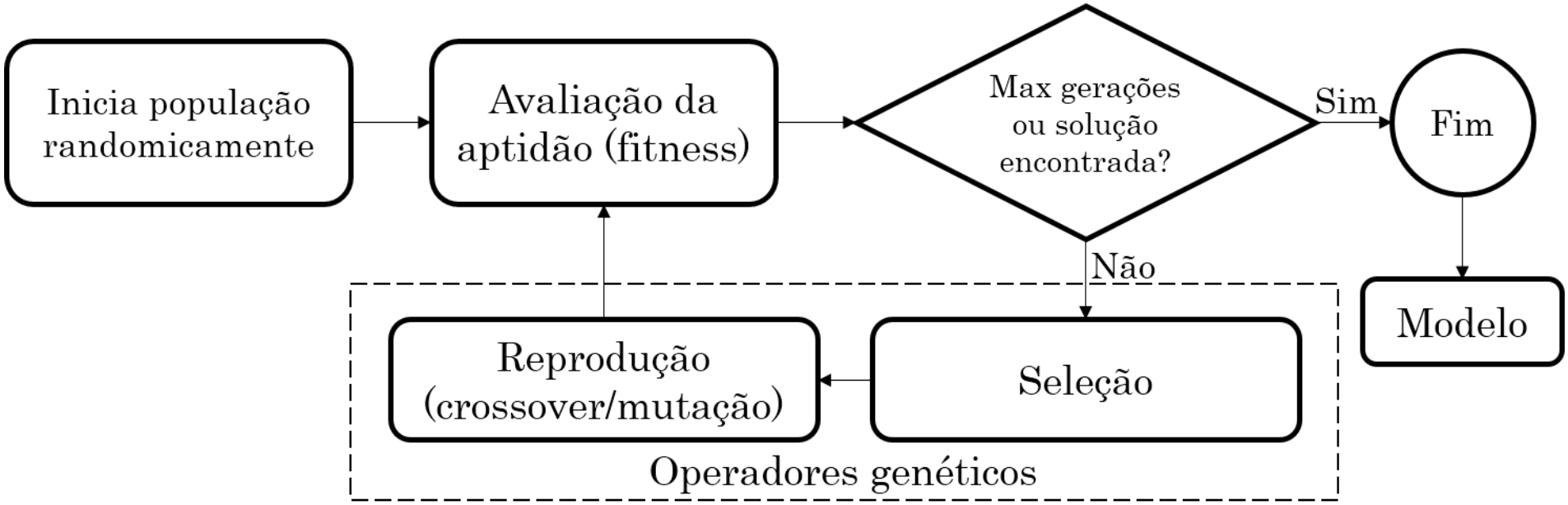


Programação Genética

- Modelos geralmente representados por meio de árvores



Programação Genética



Roteiro

- Introdução
  - Contextualização
  - Problema
  - Objetivo
- Conceitos
- Materiais e Métodos
- Análise dos Resultados
- Conclusão

Materiais e Métodos

1

TRAIN DATA

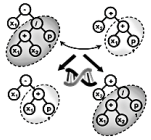
2

GENETIC PROGRAMMING

LEARNING PROCESS

GENETIC OPERATORS

MATE
MUTATION
SELECTION



FUNCTIONS

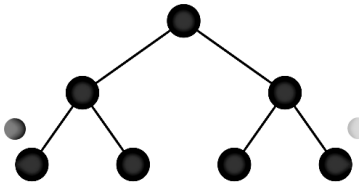
STEM, LEMMA
PONCTUACTION
POLARITY SUM
EMOTICONS, HASHTAGS
...

DICTIONARIES

POS/NEG WORDS
EMOTICONS
SLANGS
...

3

MODEL

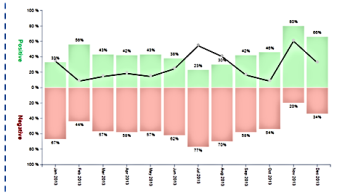


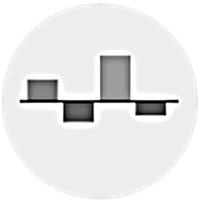
4

INPUT MESSAGES

5

CLASSIFICATION





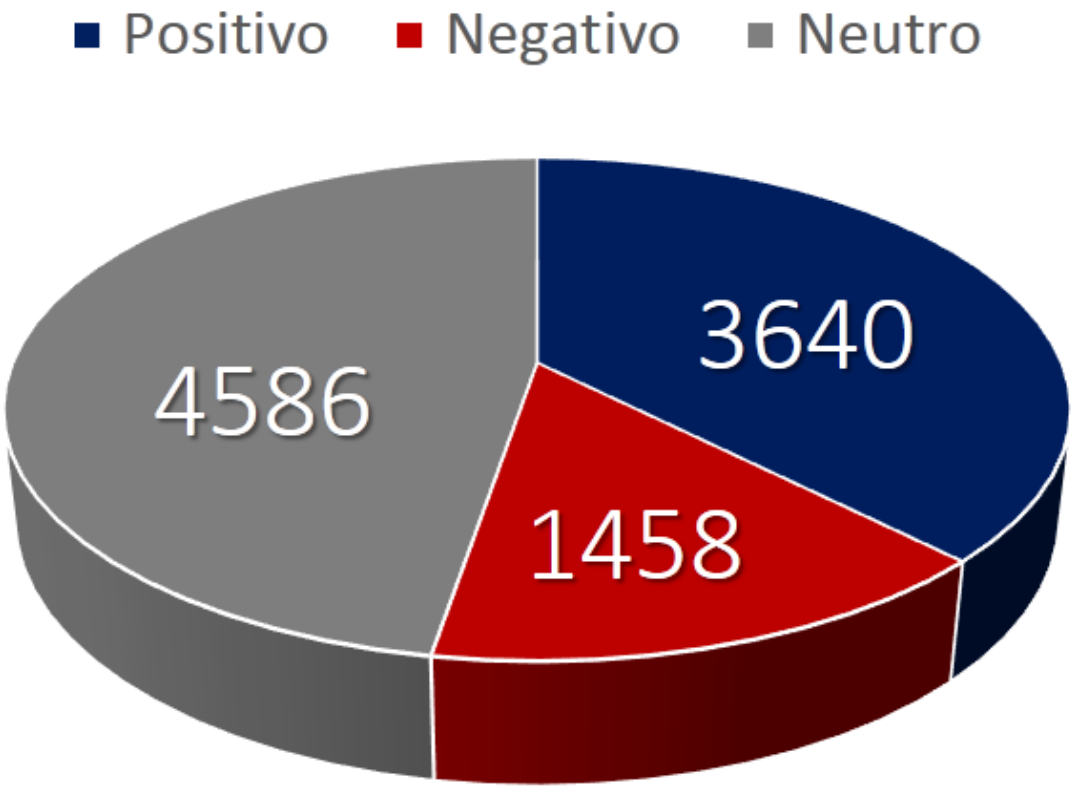
positive

42% positive
44% neutral
14% negative

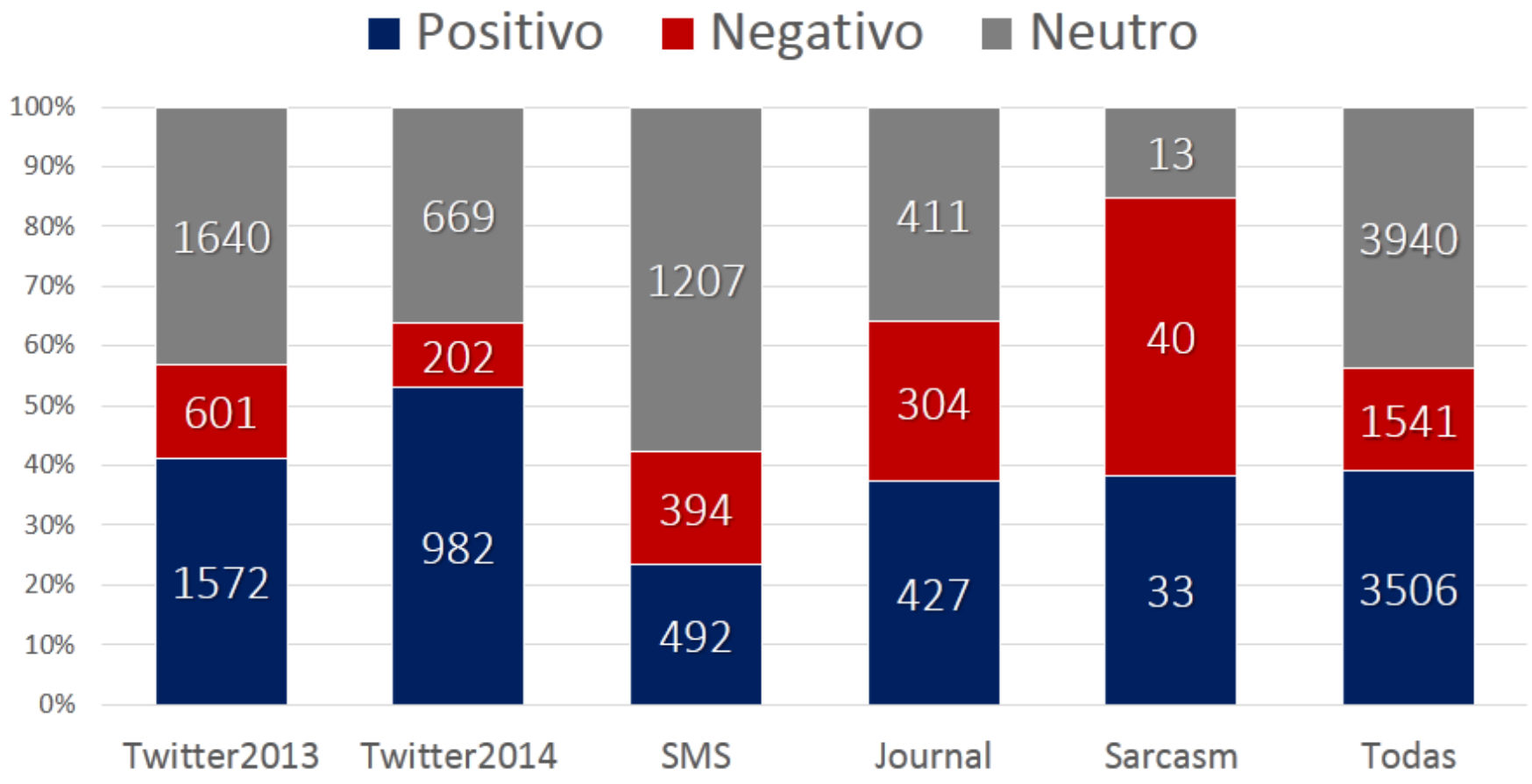
Materiais e Métodos

- *Benchmark* utilizado: SemEval 2014
  - Base de treinamento
  - Base de teste
  - Ranking dos trabalhos
- Dicionários
  - *Opinion Lexicon*
    - **2006** palavras positivas, **4800** palavras negativas
  - Emoticons
    - **186** emoticons positivos, **166** emoticons negativos

Materiais e Métodos – Base treino



Materiais e Métodos – Base teste



Materiais e Métodos

- Parametrização Programação Genética

Modelo	População	Gerações	Crossover	Mutação
A	50	500	35%	15%
B	50	600	95%	35%
C	100	650	45%	25%

Materiais e Métodos

- Funções Programação Genética (20)
  - positiveHashtags
  - negativeHashtags
  - positiveEmoticons
  - negativeEmoticons
  - polaritySum
  - hashtagPolaritySum
  - emoticonsPolaritySum
  - positiveWords
  - negativeWords
  - hasHashtag
  - hasEmoticons
  - if_then_else
  - stemmingText
  - removeStopWords
  - removeLinks
  - removeEllipsis
  - removeAllPonctuation
  - replaceNegatingWords
  - replaceBoosterWords
  - boostUpperCase

Materiais e Métodos - *Baseline*

- Modelo simples criado para comparação com os modelos gerados
- Soma simples das polaridades de cada palavra da frase
- Muito utilizado em classificadores criados manualmente

I	love	this	camera
0	+1	0	0

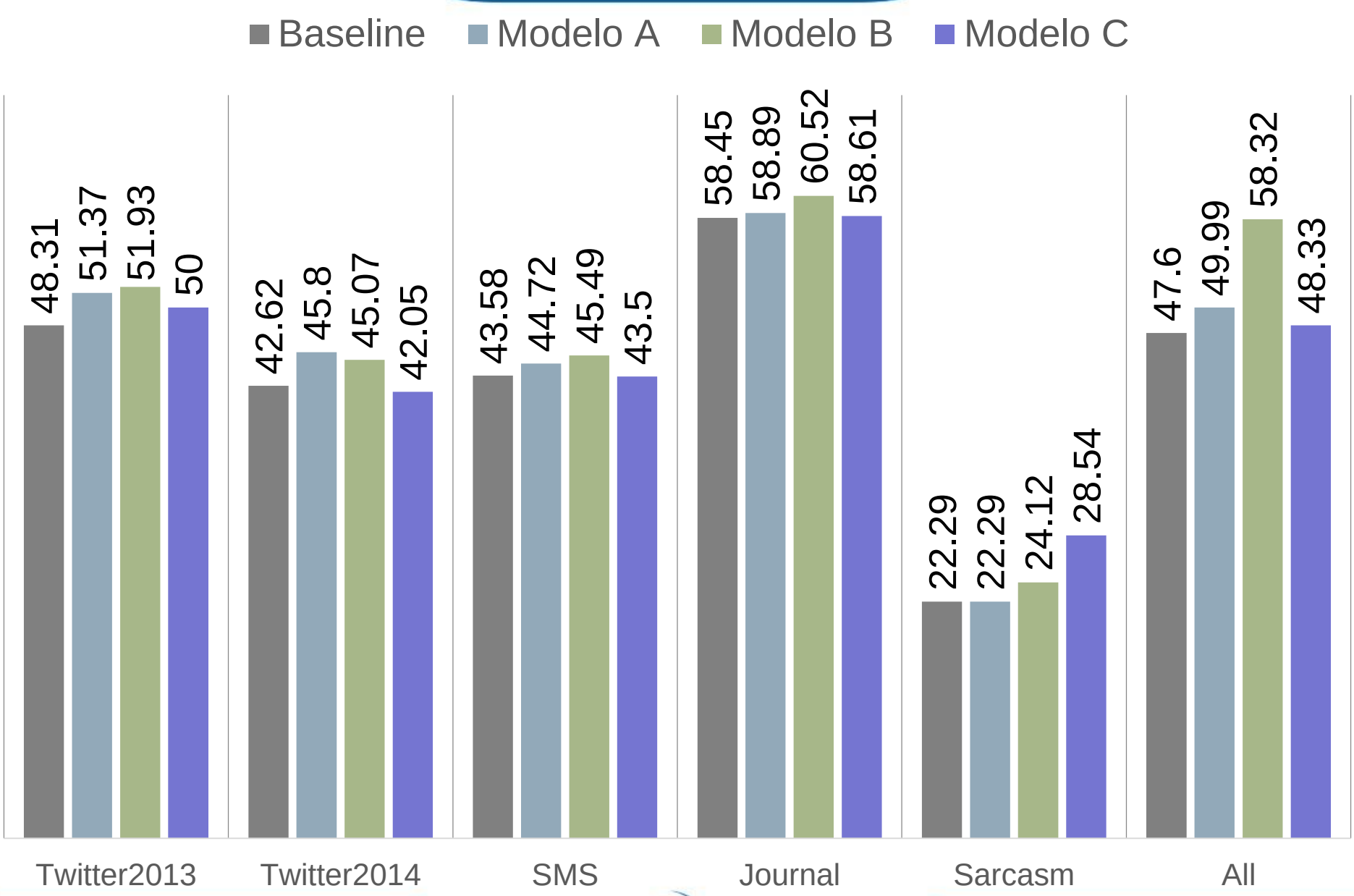
=

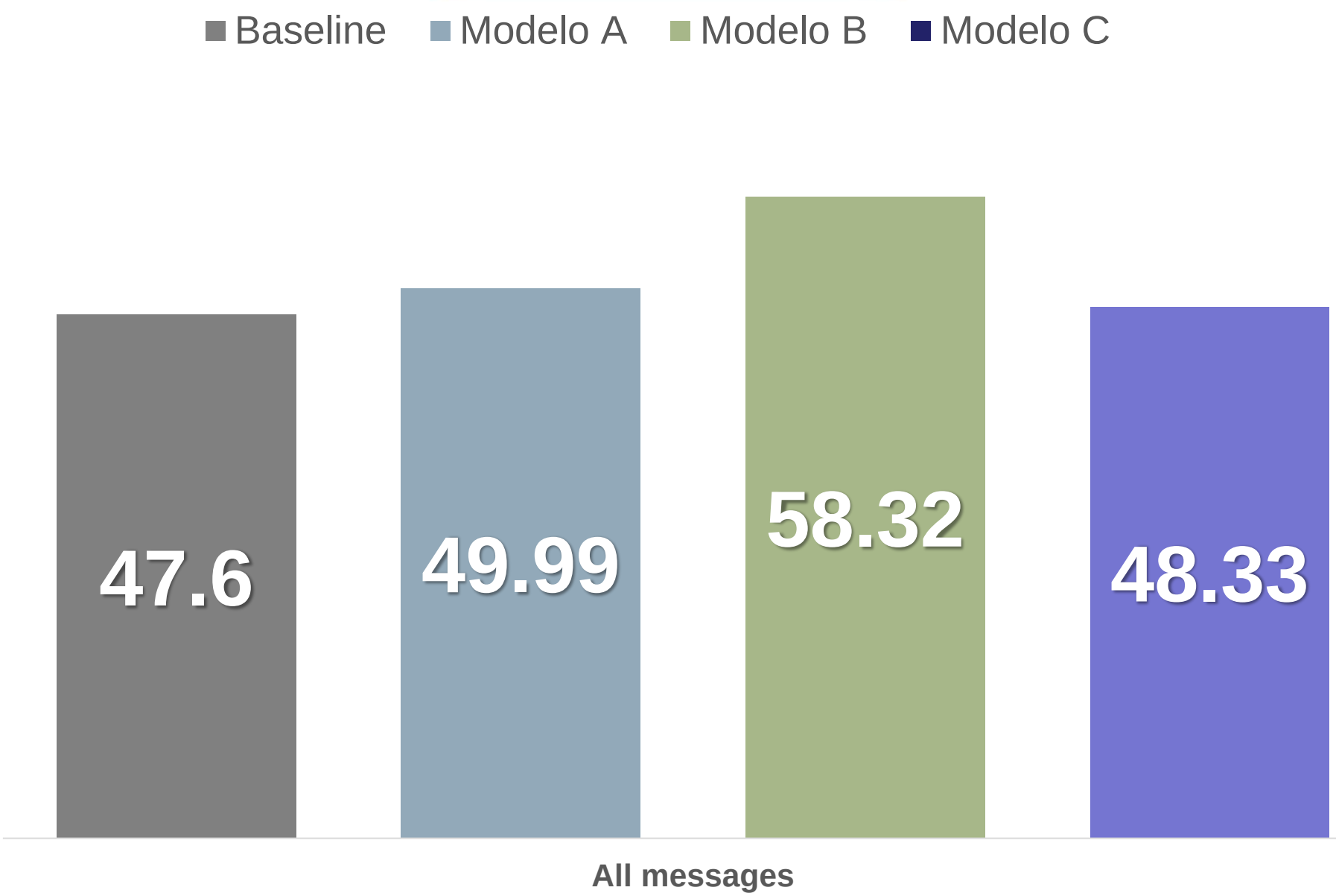
Roteiro

- Introdução
  - Contextualização
  - Problema
  - Objetivo
- Conceitos
- Materiais e Métodos
- Análise dos Resultados
- Conclusão

Materiais e Métodos - *Baseline*

- Métricas mais utilizadas para avaliação dos modelos
  - Acurácia
  - Precisão
  - *Recall*
  - F1
- Métrica utilizada: F1
  - Média harmônica de Precisão e *Recall*


Baseline
Modelo A
Modelo B
Modelo C
48.31
51.37
51.93
50
42.62
45.8
45.07
42.05
43.58
44.72
45.49
43.5
58.45
58.89
60.52
58.61
22.29
22.29
24.12
28.54
47.6
49.99
58.32
48.33
Twitter2013
Twitter2014
SMS
Journal
Sarcasm
All


Baseline
Modelo A
Modelo B
Modelo C
47.6
49.99
58.32
48.33
All messages

Resultados

- Em relação ao *baseline*

Base teste	Modelo A	Modelo B	Modelo C
Twitter2013	+6%	+4%	+4%
Twitter2014	+7%	-1%	-1%
Sarcasm	0%	+8%	**+12%**
SMS2013	+3%	+5%	0%
LiveJournal	+1%	+4%	+1%
Todas	+5%	**+6%**	+2%

Roteiro

- Introdução
  - Contextualização
  - Problema
  - Objetivo
- Conceitos
- Materiais e Métodos
- Análise dos Resultados
- Conclusão

Conclusão

- Alguns modelos apresentaram melhores resultados em determinadas sub-bases de teste
- Em todas as bases o F1 médio dos modelos gerados pela PG foram superiores ao *baseline*

Trabalhos futuros

- Melhorar inicialização da população
- Incluir novas funções para uso da Programação Genética
- Testar novas combinações de parâmetros do algoritmo
- Ampliar conjunto de dicionários
- Ampliar base de treinamento

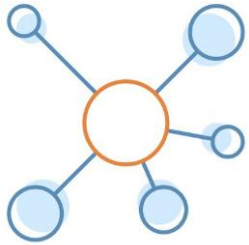
Aplicando Programação Genética na Geração de Classificadores de Sentimento

Airton Bordin Junior
airtonbjunior@gmail.com

Prof. Dr. Nádia Félix Felipe da Silva
nadia@inf.ufg.br

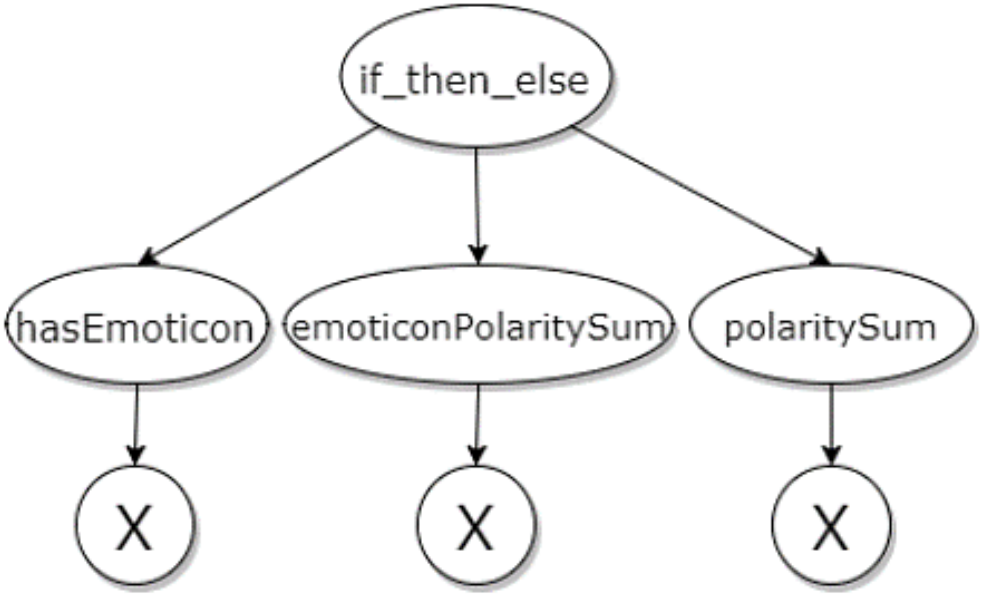
Prof. Dr. Celso Gonçalves Camilo Junior
celso@inf.ufg.br

Prof. Dr. Thierson Couto Rosa
thierson@inf.ufg.br

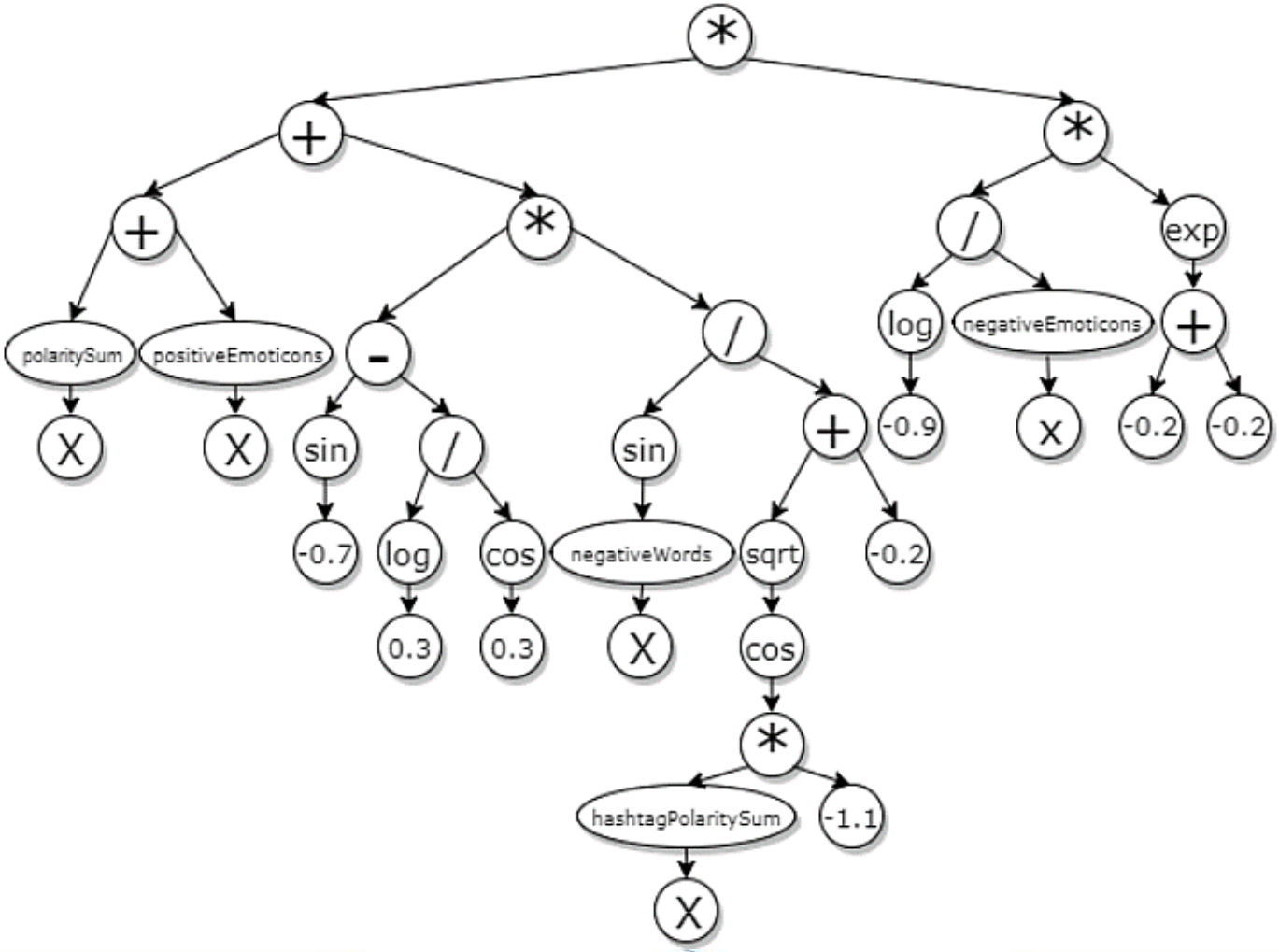


CBIC 2017

Resultados – Modelo A



Resultados – Modelo B



Resultados – Modelo C