**Projeto de TAIA – Informações da Etapa de Processamento e Avaliação**

**Dados da Equipe**

|  |  |  |
| --- | --- | --- |
| Nome | Idade | Formação |
| André Felipe Gouveia Farias | 22 | Aluno de Graduação |
| Vinícius de Oliveira Andrade | 22 | Aluno de Graduação |

**Processamento**

* Corpus Balanceado (descrever as classes utilizadas no corpus)

|  |  |  |
| --- | --- | --- |
|  | Positivo | Negativo |
| Quantidade | 201 | 201 |

* Corpus Desbalanceado (descrever as classes utilizadas no corpus)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Positivo | Negativo | Ambas(Positivo e Negativo) | Neutra | Irrelevante | Ironia |
| Quantidade | 414 | 201 | 9 | 122 | 41 | 34 |

* Resultado Corpus Balanceado (Accuracy)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Id | Pré-processamento | Naive Bayes | Support Vector Machine | Regressão Linear |
| 1 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos e dos emog  Stemização  Bag-of-Word + Frequencia | 79,12% | 79,11% | 81,34% |
| 2 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos e dos emog  Stemização  Bag-of-Word + TF-IDF | 78,11% | 80,85% | 80,35% |
| 3 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos e dos emog  Stemização  Bag-of-Word + Frequencia + Word-Bigram | 80,60% | 78,85% | 80,11% |

* Resultado Corpus Desbalanceado (Accuracy)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Id | Pré-processamento | Naive Bayes | Support Vector Machine | Regressão Linear |
| 1 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos e dos emog  Stemização  Bag-of-Word + Frequencia | 81,47% | 78,05% | 79,19% |
| 2 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos e dos emog  Stemização  Bag-of-Word + TF-IDF | 73382% | 80,49% | 75,44% |
| 3 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos e dos emog  Stemização  Bag-of-Word + Frequencia + Word-Bigram | 80397% | 75,62% | 78,53% |

* Resultado Corpus Balanceado versão 2 (Accuracy)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Id | Pré-processamento | Naive Bayes | Support Vector Machine | Regressão Linear |
| 1 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emog e dos acentos  Stemização  Bag-of-Word + Frequencia | 81,95% | 81,63% | 79,68% |
| 2 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emog e dos acentos  Stemização  Bag-of-Word + TF-IDF | 81,13% | 80,65% | 80,32% |
| 3 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emog e dos acentos  Stemização  Bag-of-Word + Frequencia + Word-Bigram | 81,30% | 80,49% | 80,97% |

O “Resultado Corpus Balanceado versão 2” é o resultado da junção do pré-processamento de ambos os membros da equipe, utilizando o mesmo corpus balanceado com duas classes que são positivo e negativo.