**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, dos emogs e dos acentos  Bag-of-Word + Frequencia | 78.86% | 78.37% | 80.61% |
| 2 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emogs e dos acentos  Bag-of-Word + TF-IDF | 79.37% | 82.09% | 82.34% |
| 3 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emogs e dos acentos  Bag-of-Word + Frequencia + Word-Bigram | 75.87% | 78.62% | 78.88% |

* 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, dos emogs e dos acentos  Bag-of-Word + Frequencia | 61.39% | 57.98% | 59.44% |
| 2 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emogs e dos acentos  Bag-of-Word + TF-IDF | 53.95% | 61.14% | 58.59% |
| 3 | Tokenização  Remoção das stopwords, das URL, das hastags, das mentions, dos nomes dos candidatos, dos emogs e dos acentos  Bag-of-Word + Frequencia + Word-Bigram | 60.66% | 58.47% | 60.41% |