

P.PORTO

**ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO**

Machine Learning

TRABALHO PRÁTICO - AVALIAÇÃO CONTÍNUA

MESTRADO EM ENGENHARIA INFORMÁTICA

HUGO GUIMARÃES – 8220337

DIOGO PEREIRA – 8200594

11 DE DEZEMBRO DE 2025

Conteúdo

1	Introdução	3
2	Engenharia de Atributos (Feature Engineering)	4
2.1	Visão Geral da Estratégia	4
2.2	Pré-processamento e Limpeza de Dados	4
2.2.1	Limpeza de Ruído	4
2.2.2	Normalização de Texto	4
2.2.3	Tratamento de Stopwords	4
2.2.4	Lematização e Stemming	4
2.3	Extração de Features Linguísticas e Estilísticas	4
2.3.1	Complexidade Lexical	4
2.3.2	Padrões de Pontuação e Formatação	4
2.3.3	Análise de Erros Gramaticais	4
2.4	Extração de Features Semânticas (NLP)	5
2.4.1	Análise de Sentimento e Subjetividade	5
2.4.2	Vetorização de Texto (Representação Numérica)	5
2.5	Extração de Features de Contexto e Metadados (Pode não ser aplicável)	5
2.5.1	Análise da Fonte / URL	5
2.5.2	Informações Temporais	5
2.5.3	Propagação Social	5
2.6	Seleção e Otimização de Features	5
2.6.1	Análise de Correlação	5
2.6.2	Redução de Dimensionalidade	5
2.6.3	Ranking de Importância	5
2.7	Matriz de Features Final	6
3	Conclusão	7

Índice de Figuras

1 Introdução

2 Engenharia de Atributos (Feature Engineering)

2.1 Visão Geral da Estratégia

[Breve parágrafo introduzindo a abordagem geral: se focaste mais em conteúdo semântico, metadados ou estilo de escrita.]

2.2 Pré-processamento e Limpeza de Dados

2.2.1 Limpeza de Ruído

[Remoção de URLs, tags HTML, caracteres especiais e emojis.]

2.2.2 Normalização de Texto

[Conversão para minúsculas (lowercase), remoção de acentuação.]

2.2.3 Tratamento de Stopwords

[Lista de palavras removidas e justificativa.]

2.2.4 Lematização e Stemming

[Técnica escolhida para reduzir as palavras à sua raiz.]

2.3 Extração de Features Linguísticas e Estilísticas

Nota: Fake news tendem a usar linguagem mais agressiva, erros e excesso de formatação.

2.3.1 Complexidade Lexical

[Contagem de palavras, tamanho médio das frases, diversidade de vocabulário.]

2.3.2 Padrões de Pontuação e Formatação

[Frequência de letras maiúsculas (ALL CAPS), uso excessivo de pontos de exclamação (!!!).]

2.3.3 Análise de Erros Gramaticais

[Se aplicável, uso de ferramentas para contar erros de sintaxe.]

2.4 Extração de Features Semânticas (NLP)

2.4.1 Análise de Sentimento e Subjetividade

[Extração de polaridade (positivo/negativo) e subjetividade (fato vs. opinião).]

2.4.2 Vetorização de Texto (Representação Numérica)

Opção A: Bag of Words (BoW) ou N-Grams.

Opção B: TF-IDF (Term Frequency-Inverse Document Frequency).

Opção C: Word Embeddings (Word2Vec, GloVe ou BERT) – especifique qual usou.

2.5 Extração de Features de Contexto e Metadados (Pode não ser aplicável)

2.5.1 Análise da Fonte / URL

[Idade do domínio, presença em listas negras, TLD (.com vs .xyz).]

2.5.2 Informações Temporais

[Data da publicação vs. Data do evento relatado.]

2.5.3 Propagação Social

[Número de partilhas, likes ou comentários (se o dataset tiver).]

2.6 Seleção e Otimização de Features

2.6.1 Análise de Correlação

[Matriz de correlação para remover variáveis redundantes.]

2.6.2 Redução de Dimensionalidade

[Uso de PCA (Principal Component Analysis) ou SVD, se houver muitos dados.]

2.6.3 Ranking de Importância

[Métodos estatísticos (Chi-quadrado, ANOVA) ou baseados em modelos (Feature Importance de Random Forest) para escolher as melhores.]

2.7 Matriz de Features Final

[Resumo final: Quantas colunas/variáveis restaram para entrar no modelo de treino e descrição do formato final dos dados (X_train).]

3 Conclusão