UNIVERSIDADE PAULISTA

HENRIQUE RODRIGUES

MATHEUS ISSAO NUMATA

**UM ESTUDO SOBRE ALGORITMOS DE CLASSIFICAÇÃO PARA ANÁLISE DE SENTIMENTOS DE TWEETS EM PORTUGUÊS**

SÃO JOSÉ DO RIO PRETO

2020

HENRIQUE RODRIGUES

MATHEUS ISSAO NUMATA

**UM ESTUDO SOBRE ALGORITMOS DE CLASSIFICAÇÃO PARA ANÁLISE DE SENTIMENTOS DE TWEET EM PORTUGUÊS**

Trabalho de conclusão de curso para obtenção do título de graduação em Ciência da Computação apresentado à Universidade Paulista – UNIP.

Orientador: xxx

SÃO JOSÉ DO RIO PRETO

2020

HENRIQUE RODRIGUES

MATHEUS ISSAO NUMATA

**UM ESTUDO SOBRE ALGORITMOS DE CLASSIFICAÇÃO PARA ANÁLISE DE SENTIMENTOS DE TWEET EM PORTUGUÊS**

Trabalho de conclusão de curso para obtenção do título de graduação em Ciência da Computação apresentado à Universidade Paulista – UNIP.

Aprovado em:

Banca Examinadora

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_/\_\_

Prof. X

Universidade Paulista – UNIP

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_/\_\_

Prof. Y

Universidade Paulista - UNIP

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_/\_\_

Prof. Z

Universidade Paulista - UNIP

**DEDICATÓRIA**

**AGRADECIMENTOS**

**EPÍGRAFE**

**RESUMO**

Palavras-chave:.

**ABSTRACT**

Keywords:.

**LISTA DE ILUSTRAÇÕES**

**LISTA DE TABELAS**

**LISTA DE ABREVIATURAS E SIGLAS**

**LISTA DE SÍMBOLOS**

**SUMÁRIO**

*Correção do Sumário*  
*Ordem e prioridade que devemos dar para os tópicos*

*A API utilizada na versão gratuita traz apenas os dados de 7 dias atrás ao realizar uma busca, devido à essa limitação a base de dados para treinamento não é muito extensa.*

1. INTRODUÇÃO
   1. TEMA E PROBLEMA DE PESQUISA
   2. JUSTIFICATIVA
   3. OBJETIVOS
      1. OBJETIVO GERAL
      2. OBJETIVOS ESPECÍFICOS
   4. METODOLOGIA DO ESTUDO
   5. ESTRUTURA DO ESTUDO
   6. TRABALHOS ANTERIORES

*Como é feita a avaliação referente a citações?*

1. FUNDAMENTAÇÃO TEÓRICA
   1. DATA SCIENCE
   2. MACHINE LEARNING
   3. CLASSIFICADORES
      1. NAIVE BAYES
      2. SUPPORT VECTOR MACHINE
   4. PRÉ-PROCESSAMENTO
      1. DATA SET (*Proporções)*
      2. STOP WORDS
      3. STEMMING
      4. XXX
2. DESENVOLVIMENTO DO PROJETO
   1. MODELAGEM
      1. CLASSIFICAÇÃO DOS TWEETS
   2. TÉCNICAS UTILIZADAS
      1. TOKENIZAÇÃO
      2. STOP WORDS
      3. XXX
   3. ALGORITMOS
      1. NAIVE BAYES
      2. SUPPORT VECTOR MACHINE
   4. CODIFICAÇÃO

*Tipos de testes*

1. TESTES
   1. ANÁLISES
   2. RESULTADOS
      1. NAIVE BAYES
      2. SVM
      3. CONJUNTO
2. CONCLUSÃO
3. REFERÊNCIAS BILIOGRÁFICAS

**1. INTRODUÇÃO**

**1.1 TEMA E PROBLEMA DE PESQUISA**

**1.2 OBJETIVOS**

**1.2.1 OBJETIVO GERAL**

Realizar uma análise comparativa dos algoritmos classificadores Naive Bayes e Support Vector Machine visando identificar discursos de ódio no Twitter.

**1.2.2 OBJETIVOS ESPECÍFICOS**

**1.3 METODOLOGIA**

**1.4 ESTRUTURA DO ESTUDO**

**1.5 TRABALHOS ANTERIORES**

**2. FUNDAMENTAÇÃO TEÓRICA**

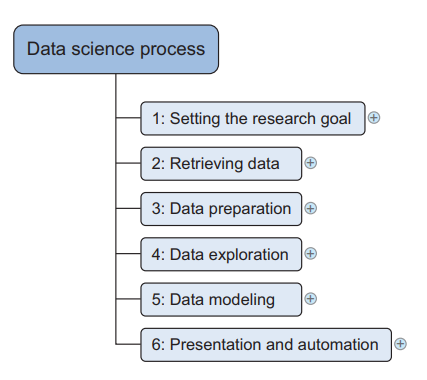
**2.1 CLASSIFICAÇÃO**

**2.1.1 DATA SCIENCE**

A Ciência de dados é um campo relativamente novo embora seus componentes já tenham sido estudados na área da computação. Sua integração acontece através de álgebra Linear, modelagem estatística, visualização, linguagem computacional, análise de gráficos, aprendizado de máquina, inteligência de negócios, armazenamento e recuperação de dados (BOSCHETTI, MASSARON, 2015).

O processo de *data science* geralmente consiste de seis passos, conforme a figura 1.1.

*Figura 1.1 – Processos do Data Science*



Fonte: <http://bedford-computing.co.uk/learning/wp-content/uploads/2016/09/introducing-data-science-machine-learning-python.pdf>

**2.1.1.1 DEFININDO O OBJETIVO DE PESQUISA**

A primeira etapa desse processo é definir o objetivo de pesquisa. O principal objetivo nesta etapa é definir o que é o projeto, com ele será feito e qual o motivo.

**2.1.1.2 COLETA DE DADOS**

A segunda etapa é a coleta de dados. É necessário utilizar dados disponíveis para análise, portanto, precisa-se encontrar dados adequados e conseguir acesso legal aos dados do dono. O resultado é coletado de forma bruta, que necessitam de um tratamento para que estes dados se tornem utilizáveis.

**2.1.1.2 PREPARAÇÃO DOS DADOS**

Após a coleta dos dados, é necessário realizar o tratamento deles. O tratamento inclui transformar os dados de um formato bruto para em um formato possível de ser utilizado pelo modelo a ser treinado.

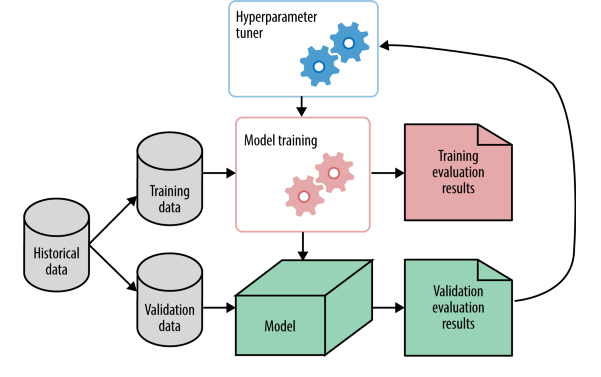
**2.1.1.3 EXPLORAÇÃO DOS DADOS**

A exploração de dados consiste em obter uma visão mais profunda a respeito dos dados. Padrões, correlações e divergências devem ser observadas nestas etapas. As observações feitas permitem a transição para a etapa de modelagem

**2.1.1.4 MONTAGEM DO MODELO**

Nesta etapa ocorre a utilização do conhecimento referente aos dados encontrados e trabalhados nas etapas anteriores. Com base nos dados encontrados, é feito a escolha do modelo a ser utilizado.

*Figura x.x – Construindo o modelo*



Fonte: <https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492>

A modelagem é dividida em três estágios distintos – treino, validação e teste. Os estágios podem sofrer alterações em caso de o modelo de aprendizado do modelo ser não-supervisionado. O final da modelagem é caracterizado pela avaliação final do modelo.

**2.1.1.5 VISUALIZAÇÃO DE RESULTADOS**

Interpretar e visualizar os dados refere-se à apresentação dos dados de uma forma não técnica. Os resultados é a forma de obtenção de um feedback à respeito do modelo utilizado.

**2.1.2 MACHINE LEARNING**

Segundo NILSSON, NILS, O machine learning geralmente se refere às mudanças em sistemas que realizam tarefas associadas à inteligência artificial (IA). Tais tarefas envolvem reconhecimento, diagnóstico, planejamento, controle do robô, previsão, etc. As “mudanças” podem ser melhorias para sistemas já em execução ou iniciação de novos sistemas.

*Machine learning* pode aparecer de várias maneiras. Grande parte do objetivo do aprendizado de máquina é reduzir uma gama de problemas bastante díspares a um conjunto de protótipos bastante estreitos. Muito da ciência do aprendizado de máquina é, então, resolver esses problemas e fornecer boas garantias para as soluções

**2.1.3 NAIVE BAYES**

**NAIVE BAYES**

Segundo JOHN, LANGLEY, 1995, o Naive Bayes é um algoritmo de classificação probabilístico, baseado na aplicação do teorema de Bayes para determinar a classe de maior probabilidade para cada instância a ser classificada. Conhecida como “ingênuo”, o Naive Bayes desconsidera completamente a correlação entre variáveis (features). Ou seja, trata cada feature de forma independente.

Após treinar um classificador Bayesiano, dado uma amostra ele decide pela classe com maior probabilidade, representado por . Essa probabilidade é calculada pela equação 1.

*Figura 1 – Equação de cálculo de probabilidade*

Fonte: <https://www.sciencedirect.com/science/article/pii/B9781558603325500559>

O algoritmo Naive Bayes considera que os atributos que descrevem as ins treinar um classificador Bayesiano, dado uma amostra ele decide pela classe com maior probabilidade, representado por .

**TEOREMA DE BAYES**

A proposta da inferência Bayesiana é conhecida desde o trabalho de Bayes e foi aplicada pela primeira vez em classificação de texto (MOSTELLER, WALLACE, 1964).

A principal suposição que caracteriza o Naive Bayes como naive (ingênuo) é a hipótese de independência das features, ou seja, ela considera que a ocorrência detodas as apalvras é independente umas das outras dado o contexto. Outra suposição é a do modelo conhecido como bag of words, ou saco de palavras, onde pressupõe-se que a ordem das palavras não possui importância, desta forma, não afeta o resultado e torna o classificador mais simples.

**NAIVE BAYES MULTINOMIAL**

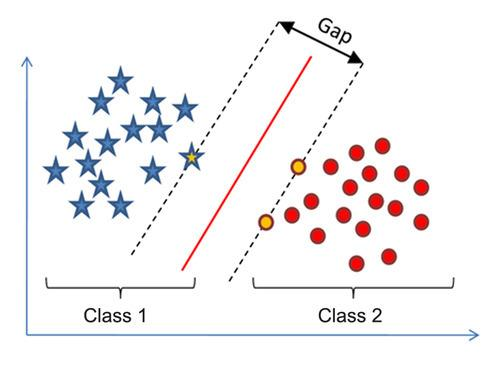
O Naive Bayes Multinomial é um algoritmo de aprendizado supervisionado bastante comum na literatura de Aprendizado de Máquina. Apesar de sua simples construção, apresenta bons resultados, sendo assim um interessante objeto de estudo. Assim como outros métodos supervisionados, este baseia seu aprendizado num modelo probabilístico. Pode-se notar que a tarefa de classificação, quando executada por humanos em documentos de texto, se dá pela observação de agrupamentos, sequências, frequências de palavras localizadas no texto. Sendo assim, podemos ver que o processo de decisão humano é intrinsecamente complexo, sendo difícil simulá-lo ou até mesmo modelá-lo devido a sua subjetividade.

Este algoritmo, no entanto, supõe uma simplificação do processo de decisão na qual as palavras e suas frequências são o único fator decisivo na determinação da classe de um documento. Essa simplificação, apesar de grosseira, facilita a modelagem matemática do problema e fornece relativamente bons resultados. O problema em questão se trata em determinar a qual classe *ci*, dentro um conjunto de classes C, um centro documento d pertence. Este evento, convenientemente representado por *ci | d*, pode ser estimado através de sua probabilidade.

**2.1.4 SUPPORT VECTOR MACHINE**

As SVMs ou Máquinas de Vetores de Suporte (SVM – do inglês, *Support Vector Machine*) são técnicas de aprendizado de máquina que utilizam o tipo de aprendizado supervisionado, um algoritmo de SVM pode ser utilizado para tarefas que abordam classificação ou regressão de dados. O treinamento é baseado na representação de cada objeto como um ponto em um espaço *n*-dimensional, onde *n* é o número atributos e com um valor de uma determinada coordenada sendo atribuído a cada atributo (BOSER, GUYON e VAPNIK, 1992).

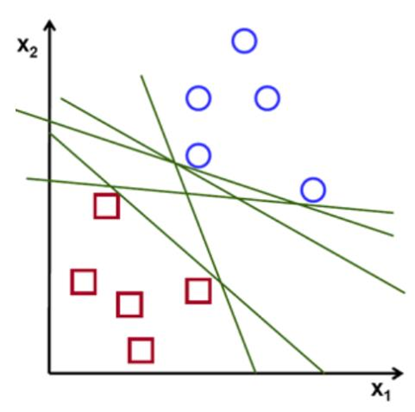
*Figura 1: Máquina de Vetores de Suporte classificação binária*



Fonte: <https://www.quora.com/Could-someone-explain-this-joke-What-did-one-support-vector-say-to-another-I-feel-so-marginalized>

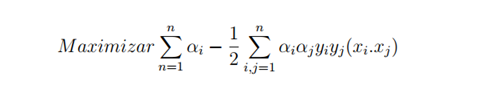
Esse trabalho forcará apenas no SVM linear, onde demonstra resultados satisfatórios na classificação de textos (YANG; LIU et al., 1999). Para ajudar na compreensão do SVM linear uma outra descrição exemplifica a classificação, um SVM tem como objetivo separar um conjunto de dados por exemplo classificados como coordenadas X ou Y, por uma linha reta (um hiperplano) em um espaço dimensional de tamanho *n,* onde *n* é a quantidade total de características (*features*), conforme (BURGES,1998).

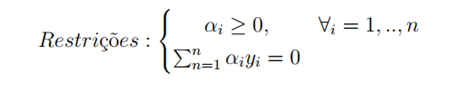
*Figura 2- Objetos separados por hiperplanos*



Fonte: (DRAKOS,2018)

A Figura 2 representa objetos diferentes separados por vários hiperplanos de forma aleatória, o SVM maximiza o tamanho da margem entre as separações dos hiperplanos. Para compreender essa separação suponha um conjunto de dados para treinamento {xi...xn} e suas respectivas classificações {yi...yn} {-1,1}, encontrar a maior margem dentre os separadores do hiperplano é um problema de otimização, formulado como (LORENA; CARVALHO,2007):

(Eq.1)

(Eq.2)

Onde α*i* é um parâmetro que varia de *i* até *n*.

**2.2 PRÉ-PROCESSAMENTO**

**2.2.1 DATASET**

Nós começamos com um dicionário léxico de discurso de ódio contendo palavras e frases identificadas pelos usuários como ofensas. Utilizando a API do Twitter, buscamos por tweets contendo termos do dicionário léxico, resultando em 1500 amostras de tweets. Dessas amostras nós utilizamos todos os tweets e codificamos manualmente. Cada tweet foi classificado em uma das três categorias: positivo, neutro ou negativo. A classificação foi feita baseado nas palavras e no contexto em que cada tweet foi utilizado.

**2.2.2 STOP WORDS**

**2.2.3 STEMMING**

**2.2.4 XXX**

**3. DESENVOLVIMENTO**

**3.1 CLASSIFICAÇÃO DOS TWEETS**

**3.2 PRÉ-PROCESSAMENTO**

**3.2.1 TOKENIZAÇÃO**

**3.2.2 STOP WORDS**

**3.2.3 STEMMING**

**3.3 NAIVE BAYES**

**3.4 SVM**

**4. TESTES E RESULTADOS**

**4.1 NAIVE BAYES**

**4.2 SVM**

**5. DISCUSSÃO**

**6. CONCLUSÃO**

**7. REFERÊNCIAS**

GOOD, I. J. 1965. The Estimation of Probabilities: An Essay on Modern Bayesian Methods. M.I.T. Press

NILSSON, N. J. 1996. Introduction to Machine Learning: An early draft of a proposed textbook. Stanford University

JOHN, G. H. 1995. UAI’95: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence

MOSTELLER, F.; WALLACE, L. D.; Applied Bayesian and Classical Inference: The Case of the Federalist Papers. New York: Springer New York, 1984.

FRANK, E. BOUCKAERT, R. R. 2006 Naive Bayes for Text Classification with Unbalanced Classes

CORTES, C. e VAPNIK, V., “Support Vector Networks”, Machine Learning, v.20, p.273-297, 1995

VAPNIK, V., “The Nature of Statistical Learning Theory”, Springer 1995

YANG, Y. e PEDERSEN, J. O.,“Comparative Study on Feature Selection in Text Categorization”, Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997

CHIH-WEI Hsu e CHIH-JEN LIN, “A Comparison of Methods for Multi-class Support Vector Machines”, Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2001.

BURGES, C. J. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, Springer, v. 2, n. 2, p. 121–167, 1998.

YANG, Y.; LIU, X. et al. A re-examination of text categorization methods. In: Sigir. [S.l.: s.n.],1999. v. 99, n. 8, p. 99.

BOSCHETTI, A.; MASSARON, L. Python Data Science Essentials, Birmingham 2015.