UNIVERSIDADE FEDERAL DO ACRE
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

AVALIAÇÃO DE DESEMPENHO DE ALGORITMOS DE CLASSIFICAÇÃO EM MINERAÇÃO DE OPINIÃO EM TEXTOS EM PORTUGUÊS

RIO BRANCO
2019

JARDEL DA CUNHA NASCIMENTO

AVALIAÇÃO DE DESEMPENHO DE ALGORITMOS DE CLASSIFICAÇÃO EM MINERAÇÃO DE OPINIÃO EM TEXTOS EM PORTUGUÊS

Monografia apresentada como exigência final para obtenção do grau de bacharel em Sistemas de Informação da Universidade Federal do Acre.

Prof. Orientador: Daricélio Moreira Soares, Dr.

RIO BRANCO

2019

TERMO DE APROVAÇÃO

JARDEL DA CUNHA NASCIMENTO

AVALIAÇÃO DE DESEMPENHO DE ALGORITMOS DE CLASSIFICAÇÃO EM MINERAÇÃO DE OPINIÃO DE TEXTOS EM PORTUGUÊS

Esta monografia foi apresentada como trabalho de conclusão de Curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre, sendo aprovado pela banca constituída pelo professor orientador e membros abaixo mencionados.

Compuseram a banca:

Prof. Olacir Rodrigues Castro Junior, Dr.
Curso de Bacharelado em Sistemas de Informação

Prof. Catarina de Souza Costa, Dra.
Curso de Bacharelado em Sistemas de Informação

Prof. Daricélio Moreira Soares, Dr.
Curso de Bacharelado em Sistemas de Informação

Rio Branco, 14 de março de 2019

À minha mãe pelo apoio emocional e financeiro
durante toda minha trajetória de estudante.

AGRADECIMENTOS

Agradeço, primeiramente, à minha mãe, que dedicou parte de sua vida fomentando minha formação como cidadão, período em que, desde a infância, alimentou hábitos de leitura e compreensão de mundo.

Agradeço também aos amigos de curso que me deram suporte e apoio emocional durante essa jornada, alguns laços de amizade, tenho certeza, irão se perpetuar por nossa existência.

"A dúvida é o princípio da sabedoria"

(Aristóteles)

RESUMO

A capacidade de processar e determinar computacionalmente o sentimento expresso em texto é uma enorme vantagem competitiva para organizações, afinal, entre várias aplicações, torna-se possível processar e quantificar o sentimento expresso nos feedbacks de clientes em comentários e publicações feitas pelas redes sociais acerca de produtos, serviços e ainda outras relações comerciais. Nesse contexto, esse trabalho propõe avaliar o desempenho de três algoritmos de classificação – *Naive Bayes, Suport Vector Machine* (SVM), *Maximum Entropy* – durante a tarefa de determinação da polaridade – positiva ou negativa – de um texto, usando as métricas: acurácia, precisão e revocação. Os experimentos foram aplicados em duas bases de dados textuais, do idioma português, previamente classificadas, sendo: i) uma coleção de revisões de filmes e ii) uma coleção de tweets classificados pela abordagem de emoticons. Ao final os classificadores *Maximum Entropy* e *SVM* apresentaram os melhores desempenhos.

Palavras-chave: Análise de Sentimentos, Mineração de Opinião, Inteligência Artificial.

ABSTRACT

The ability to process and computationally determine the sentiment expressed in text is a huge competitive advantage for organizations, after all, among various applications, it becomes possible to process and quantify the sentiment expressed in customer feedbacks in comments and publications in social networks about products, services and other business relationships. In this context, this work proposes to evaluate the performance of three classification algorithms – Naive Bayes, Suport Vector Machine (SVM), Maximum Entropy – during the task of determining the polarity - positive or negative - of a text, using the metrics accuracy, precision and recall. The experiments were applied to two Portuguese textual databases, previously classified, being: i) a collection of movie reviews and ii) a collection of tweets classified by the emoticons approach. In the end, the Maximum Entropy and SVM classifiers presented the best performances.

Key-words: Sentiment Analysis, Opinion Mining, Artificial Intelligence.

LISTAS DE FIGURAS

LISTAS DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

No mundo moderno, devido à internet e a grande quantidade de sistemas computadorizados que gerenciam informações que envolvem clientes, produtos e serviços, extrair a maior quantidade possível de vantagem desses dados é uma estratégia recorrente para suporte a otimização de produtos e serviços, assim como, do próprio relacionamento negócio-cliente. Nesse contexto, surgiram tecnologias que tornam realidade a obtenção de conhecimento novo e inexplorado oriundo dessa grande massa de Informações. Exemplo disso é a Mineração de Opinião (MO) ou Análise de Sentimentos (AS) de forma automatizada, técnica essa que possibilita a obtenção de *feedback* emocional sobre produtos, serviços, eventos e até mesmo figuras públicas.

Nesse sentido, as redes sociais são um bom exemplo de domínio para aplicação de técnicas de Mineração de Opinião. O Twitter, lançado em 2006, possui um serviço de microblog, uma espécie de “SMS da internet”, onde os usuários compartilham opiniões sobre os mais diversos temas. Fato esse que torna rede social em questão local atrativo para estudos desse tipo. Outro fator motivador é que o próprio Twitter disponibiliza uma Interface de Programação de Aplicações (API, do inglês, *Application Programming Interface*), que, dentre outras possibilidades, a ferramenta oferece funções para coleta de publicações dos milhões de usuários da plataforma.

Cheong e Lee (2011) apontam técnicas de previsão de terrorismo em seu trabalho através do uso do Twitter, isso dá uma pequena noção das várias possibilidades sobre Mineração de Texto (MT) através de publicações no Twitter. Afinal, não existe uma censura sobre a opinião expressa em texto na rede social, o que poderia acontecer pela presença de um entrevistador no caso de uma pesquisa de opinião formal.

Levar em consideração o conteúdo textual deixado por clientes em comentários é uma ótima forma de obter uma avaliação institucional, exemplo de uma ótima aplicação dessa técnica é o sistema de nota para hospedagens do site de reserva de hotéis *TripAdvisor*[1]. As notas são dadas pelo sistema de estrelas, porém os usuários não possuem a possibilidade de atribuir nota pelo voto direto, mas através de seus comentários, que após a devida avaliação de sentimento, gera uma nota para o a hospedagem. Assim, o sistema leva em consideração os próprios *feedbacks* dos clientes como métrica de avaliação das hospedagens.

Dessa forma, para criar modelos robustos e eficientes de *feedbacks* emocionais de clientes, deve-se avaliar o desempenho dos algoritmos, escolhendo assim, na aplicação final do modelo, o/os algoritmos que melhor classificarem os sentimentos. Diversos são os trabalhos que se propõem a avaliar o desempenho de algoritmos, alguns destes elencados na seção de trabalhos relacionados. Por fim, para que a avaliação de desempenho ocorra, algumas métricas, comumente utilizadas, para avaliar algoritmos desse tipo foram escolhidas, assim, avaliou-se o desempenhos dos classificadores em termos de: acurácia, precisão e revocação.

1.1 PROBLEMA DA PESQUISA

As tarefas realizadas, comumente, na internet como usar as redes sociais, comprar online e responder um *post* em fóruns deixam armazenadas na web informações que, se bem utilizadas, podem ser úteis para gerar novos conhecimento

[1] https://www.tripadvisor.com.br/

em diversos domínios. Porém, estas informações – em geral, em formato de comentários e opiniões – contidos em páginas online são feitos em linguagem natural e que máquinas não conseguem compreender sem o devido tratamento desses dados. Para os computadores, esses comentários, da forma que são disponibilizados na internet, não passam de conjunto de caracteres.

Uma das aplicações do uso desses comentários para gerar novo conhecimento é mineração de opinião através do processamento de linguagem natural; Por exemplo, saber o que pensam as pessoas inseridas em um contexto de determinado produto ou serviço sempre foi de grande interesse por parte daqueles que o oferecem, pois tal informação é de grande valia, uma vez que pode ser utilizada no processo estratégico de tomada de decisão.

Nesse sentido, esse estudo teve como norteador a seguinte questão:

- Após devida avaliação de desempenho, dado um escopo de classificadores e também dois domínios de coleções de dados, quais destes algoritmos apresentariam o melhor desempenho, nesse contexto, quando aplicados a tarefa de classificação da polaridade de sentimentos em Mineração de Opinião?

1.2 OBJETIVOS DA PESQUISA

Foram tidos como elementos norteadores dessa pesquisa os objetivos que seguem, sendo os objetivos específicos, aqueles que destrincham e traçam um caminho lógico para o devido alcance do objetivo geral.

1.2.1 OBJETIVO GERAL

O objetivo geral dessa pesquisa é investigar, testar e avaliar o desempenho de classificadores na tarefa de Análise de Sentimentos/Mineração de Opinião expressa em texto.

1.2.2 OBJETIVOS ESPECÍFICOS

Elencam-se os objetivos específicos, que nada mais são do que etapas necessárias a realização dessa pesquisa, sendo eles:

- Revisar a literatura relacionada a fim de dominar conceitos fundamentais do objeto de estudo;
- Definir coleções de dados objeto de estudo;
- Efetuar o pré-processamento dos documentos de textos obtidos;
- Selecionar algoritmos de classificação e métricas para a etapa de experimentação;
- Realizar experimentações com as coleções definidas;
- Analisar e reportar os resultados dos testes.

1.3 JUSTIFICATIVA DA PESQUISA

A quantidade de dados gerados pela atividade humana na internet cresce constantemente. Uma parte significativa desses dados são opiniões sobre produtos

e/ou serviços, que ficam armazenados em *data centers* de lojas virtuais, redes sociais e fóruns. Nesse sentido, tornou-se cada vez mais comum o investimento em ciência de dados para a mineração dessas opiniões, o que traz um *feedback* automatizado sobre produtos e serviços às empresas sem a necessidade de consultorias e pesquisas de opinião.

Os cientistas de dados, como são chamados os profissionais dessa área, realizam a descoberta desse conhecimento através da aplicação de diversos algoritmos em uma coleção de dados relacionados ao domínio estudado. Porém, os diversos algoritmos disponíveis para essa atividade apresentam desempenhos diversos, tendo melhores resultados em um cenário do que em outro. Além disso, esses dados estão armazenados de forma não estruturada ou semiestruturadas, exemplo disso são os comentários e postagens em redes sociais.

Devido à natureza desse tipo de dado, a tarefa computacional de minerar opinião não é fácil, porém, segundo Pang e Lee (2008) e Liu (2012), devido ao massivo crescimento da quantidade de dados disponíveis, impulsionados pelo uso cada vez maior de mídias sociais, cada vez mais, indivíduos e instituições têm utilizado o conteúdo das redes sociais como parte do processo de tomada de decisão.

Assim sendo, levando em consideração o cenário emergente, tanto das redes sociais, como nova forma de relacionamento social, quanto da mineração de opinião em redes sociais. Tem-se ainda como motivação, a baixa produção científica de estudos dessa área emergente no curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre (UFAC).

Ainda assim, existem muitos obstáculos para essa tarefa computacional, como por exemplo, a diversidade de línguas falada por humanos, diferenças culturais que implicam nas reações expressas em texto. O que torna pouco viável um modelo multilinguístico eficaz para essa tarefa. Nessa questão, em específico, esse estudo tem como foco a língua portuguesa como objeto de experimentação, sendo esse o idioma nativo do país onde essa pesquisa foi realizada, além disso, durante a revisão de literatura, percebeu-se que grande parte dos trabalhos semelhantes, inclusive os realizados no Brasil, têm como objeto de estudo coleções de dados de texto em inglês.

1.4 MÉTODO DE PESQUISA

Segundo orientação da obra de Wazlawick (2014), metodologicamente falando, essa monografia configura-se um estudo empírico. Suas 6 (seis) etapas estão representadas no fluxograma da Figura 1. Além disso, Gil (2010) traz a definição desse tipo de pesquisa, baseado nos procedimentos técnicos utilizados, como sendo uma pesquisa experimental.

Figura 1. Etapas do estudo

Fonte: Elaboração própria

Segue abaixo a descrição do que foi feito em cada etapa:

- **Definição das coleções de dados:** escolheu-se base de dados da língua portuguesa já rotuladas; uma de tweets e outra de revisões de filmes traduzidos automaticamente da língua inglesa;
- **Seleção dos algoritmos:** foram selecionados três algoritmos classificadores, de aprendizado de máquina supervisionado, para a experimentação, eles são abordados nas seções seguintes;

- **Definição das técnicas e métricas:** duas foram as técnicas de experimentação; sendo elas: parâmetro de particionamento da base em 10% para teste e 90% para treino do algoritmo; e a validação cruzada, onde toda a base foi particionada em 10 bases de testes com garantia de que toda entrada fosse utilizada como treino e teste. Também melhor definido nas seções seguintes; como métricas utilizou-se: acurácia, precisão e revocação;

- **Experimentação:** a execução dos testes descritos acima;

- **Avaliação dos resultados e testes:** onde os resultados dos testes são discutidos.

1.5 ORGANIZAÇÃO DO ESTUDO

Além deste capítulo, esta monografia está organizada em outros quatro capítulos, além desse, temos:

- **Capítulo 2:** que contém todo o referencial teórico necessário para a compreensão da pesquisa;

- **Capítulo 3:** onde é apresentado ao leitor todo o estudo de caso realizado e seus resultados;

- **Capítulo 4:** em que são feitas as considerações finais e também citadas as recomendações para estudos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Conforme recomendado por Wazlawick (2014), neste capítulo faremos uma abordagem dos principais temas e conceitos ligados ao objeto de estudo dessa pesquisa. Deste modo, a organização do capítulo se dá pela relação dos conceitos entre si, ou seja, partindo daquilo que é mais fundamental para o que é mais específico.

2.1 BANCO DE DADOS

A vida moderna tem colocado as pessoas em frequente contato com bancos de dados e sistemas de banco de dados por meio de tarefas diárias: uma pesquisa rápida no Google retorna resultados que estão armazenados em bancos de dados conectados à rede mundial de computadores; ao acessar o *feed* de notícias de redes sociais são apresentadas fotos, publicações e anúncios publicitários, informações igualmente armazenadas em banco de dados.

Ao responder a simples pergunta “crédito ou débito?” durante uma compra usando cartão de plástico, uma consulta em banco de dados é disparada para verificar

a autenticidade da senha e saldo na conta bancária ou crédito disponível na operadora de cartão. Até mesmo a tarefa de ir ao mercado para realizar uma compra, mesmo que com dinheiro físico, o sistema de vendas do caixa atualiza a quantidade, em estoque, do produto que acaba de ser comprado. Sistemas de banco de dados rodeiam a vida dos indivíduos atualmente.

De forma genérica, Elmasri e Navathe (2011) definem que dados propriamente dito são fatos e informações conhecidas que possuem um significado implícito e que podem ser armazenadas; já uma coleção relacionada desses dados é chamada banco de dados.

Corroborando com a definição anterior, Silberschatz et al. (2012) contribuem definindo como banco de dados a coleção de dados, que inter-relacionada e associado a programas de acesso a esses dados, formam um Sistema de Gerenciamento de Banco de Dados (SGBD), que tem como principal função armazenar e recuperar informações desse banco de dados com eficiência e conveniência.

2.1.1 DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS (KDD)

Anteriormente, o contato do ser humano com dados armazenados em banco de dados se dá de forma cotidiana e invisível para o usuário. Nesse contexto, extrair conhecimento e descobrir padrões dessas bases de dados tornou-se objeto de estudo. Assim sendo, a tarefa de Descoberta de Conhecimento em Banco de Dados (KDD, do inglês *Knowleged Discovery in Databases*) é definida como um processo difícil que visa descoberta de padrões que sejam novos e tecnicamente utilizáveis e entendíveis Fayyad et al. (1996).

O presente trabalho se baseia na definição de que, segundo Elmasri e Navathe (2011), o processo de KDD possui 6 (seis) fases sequenciais, sendo:

1. Seleção dos dados: escolher o recorte específico da base de dados em que se deseja descobrir conhecimento novo;

2. Limpeza dos dados: garantir a consistência dos dados através da remoção ou correção dos dados inconsistente;

3. Enriquecimento dos dados: melhorar a qualidade dos dados com outras fontes de informações;

4. Transformação dos dados: consiste em aplicar codificações e agrupamentos aos dados, separando-os por categorias e intervalos, quando possível, processo também chamado de normatização dos dados;

5. Mineração dos dados: etapa em que os dados são processados para a descoberta de regras e padrões entre eles, existem várias técnicas que podem ser aplicadas, como regras de associação, árvore de decisão e padrões sequenciais;

6. Relatório: apresentação do conhecimento novo extraído da análise das regras e padrões descobertos na etapa anterior.

2.1.2 MINERAÇÃO DE DADOS (*DATA MINING)*

Consultas em banco de dados têm como finalidade extrair informações específicas de um contexto limitado e não-genérico, porém, às vezes, é desejável ter informações genéricas e mais amplas para auxílio na tomada de decisão. Por exemplo, suponha uma universidade que possui vários sistemas e vários banco de dados para esses sistemas, (a) um sistema salva em banco, além das informações pessoais das pessoas, os registros sobre a catraca eletrônica, com identificação de alunos e funcionários, horários de entrada e saída; (b) também há um sistema de

registro de frequência em sala de aula que guarda os mesmos tipos de registros, mas no contexto da sala de aula; (c) há ainda um outro sistema que salva em banco de dados as informações sobre a utilização do restaurante universitário, registrando informações sobre compra e venda de créditos para refeições, assiduidade nas refeições e horários das refeições.

As consultas habituais em banco de dados podem retornar com facilidade o horário em que um aluno normalmente almoça, qual a sua média de faltas por mês, quantas vezes ele utilizou o restaurante no último mês. Porém, para a gestão institucional seria interessante, por exemplo, conseguir (a) prever quantos alunos reprovarão por falta no semestre seguinte, facilitando o planejamento pedagógico e ações afirmativas para reverter esse quadro; (b) qual a demanda de utilização do restaurante universitário em determinado horário, semana ou mês, evitando o desperdício ou falta de comida; (c) quais as características dos alunos que deixam de almoçar no restaurante, possibilitando uma readequação do planejamento semestral visando a redução de impacto no orçamento com base nos alunos egressos.

Consegue-se acesso a esse tipo de conhecimento por meio da tarefa de mineração de dados. Vale ressaltar que em no exemplo não estão sendo levandas em consideração as especificidades técnicas do banco de dados hipotético – utilizando da didática, se quer apenas elucidar a aplicação e o problema que criou a necessidade da mineração de dados.

Nesse aspecto, a KDD depende da tarefa de mineração de dados, sendo a mineração etapa primordial para a descoberta de conhecimento, só a mineração de dados constitui 3 (três) das 9 (nove) etapas de KDD definidas por Fayyad, Piatetsky-Shapiro e Smyth (1996): (i) criar o conjunto de dados a serem utilizados; (ii) entender o domínio do problema; (iii) criar o conjunto de dados a serem utilizados; (iv) limpeza e pré-processamento dos dados; (v) projeção e redução dos dados; (vi) escolher a função de mineração de dados; (vii) escolher os algoritmos de mineração, (viii) minerar os dados; (ix) interpretação dos dados.

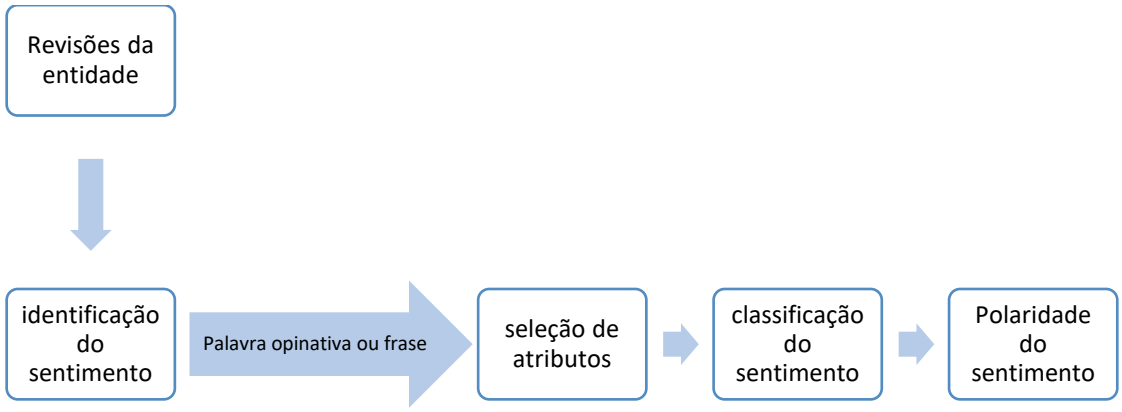
Assim sendo, Witten et al. (2011) apresenta como definição de Mineração de Dados, como sendo o processo de descoberta de padrões em dados, sendo esse, um processo automatizada ou semi-automático. Os padrões descobertos devem trazer

alguma vantagem, aplicação útil, normalmente econômica. Já, Elmasri; Navathe (2011), definem mineração de dados como sendo a “mineração ou descoberta de novas informações em termos de padrões ou regras com base em grandes quantidades de dados”.

2.2 MINERAÇÃO DE OPINIÃO

A Análise de Sentimentos (AS) ou Mineração de Opinião (MO) pode ser definida, de acordo com o exposto em Medhat et al. (2014), como o estudo computacional da opinião das pessoas a cerca de uma determinada entidade, que representa indivíduos, eventos ou tópicos. Apesar de AS e MO representarem conceitos intercambiáveis, Tsytsarau; Palpanas (2012) mostram que alguns pesquisadores podem apresentar sutis diferenciações na conceitualização dos termos, sendo a MO considerada a extração e análise da opinião das pessoas sobre uma determinada entidade, enquanto a AS identifica o sentimento expresso em texto e, em seguida, analisa.

Figura 2. Processo de análise de sentimentos

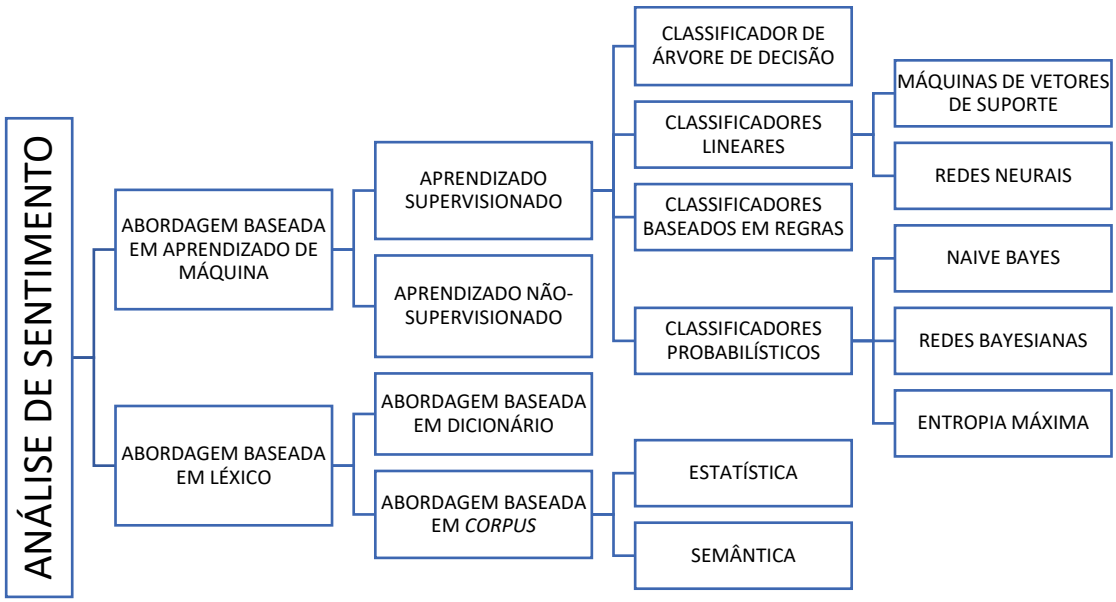


Fonte: Baseado em Medhat et al. (2014)

A Figura 2, adaptada do *survey* de Medhat et al. (2014), ilustra como a AS pode ser considerada um processo de classificação, deixando claro que o alvo da AS é

identificar opiniões, identificar os sentimentos expressos nessas opiniões e, posteriormente, classificar a polaridade deste sentimento, seja ele, positivo, negativo ou neutro. Já a Figura 3, também adaptada do mesmo *survey,* ilustra as técnicas de classificação de sentimentos e suas respectivas ramificações.

Figura 3. Técnicas de análise de sentimento



Fonte: Baseado em Medhat et al. (2014)

Liu (2012) define Análise de Sentimentos e Mineração de Opinião, como sendo o campo de estudo que analisa opiniões, sentimentos, avaliações, atitudes e emoções em relação às entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos. Em seu livro, nos são apresentados outros nomes e tarefas ligeiramente diferentes, tais como, extração de opinião, mineração de sentimentos, análise de subjetividade, análise de afeto, análise de emoção, revisão de mineração e outros; Porém, todos esses, estão embaixo do guarda-chuva que compõe a grande área de estudos de Análise de Sentimentos ou Mineração de Opinião.

2.2.1 MINERAÇÃO DE TEXTO E *WEB MINING*

Anteriormente, falou-se sobre a definição de mineração de dados, analogamente, então, conforme Witten et al. (2011), mineração de texto seria a busca por padrões em documentos de texto. Sendo a grande diferença entre mineração de dados e mineração de texto: (i) em mineração de dados trabalha-se com informações que estão implícitas nos dados base para a tarefa; (ii) já na mineração de texto a informação está explícita em texto, porém completamente escondido do ponto de vista computacional.

Witten et al. (2011) prossegue, ao afirmar que mineração de texto ainda é uma tecnologia em expansão e, por isso, ainda não há consenso sobre até onde vai a abrangência da recente área. Por fim, destaca que todo o escopo de Processamento de Linguagem Natural (PLN) se enquadra como mineração de texto.

Nesse contexto, *Web Mining* é como mineração de texto na *World Wild Web*, que por si só já se trata de um enorme repositório de texto, porém com informações extras como marcações que indicam o formato do documento ou que difere hipertextos do restante do conteúdo, ou ainda, títulos, subtítulos, tags e outros. Isso tudo, faz com que, geralmente, a mineração de texto na web, ainda segundo Witten et al. (2011), tenha melhores resultados.

2.2.2 PROCESSAMENTO DE LINGUAGEM NATURAL

Levando em consideração o que é apresentado no livro de Witten et al. (2011) e no *survey* sobre mineração de texto de Gupta e Lehal (2009), pode-se ter a definição de Processamento de Linguagem Natural (PLN) como o composto de técnicas computacionais utilizadas para a compreensão, computacionalmente falando, da linguagem natural humana, expressa em texto. Algumas dessas técnicas utilizadas nesse estudo são apresentadas no capítulo do estudo de caso.

Aranha et al. (2006) afirma ainda que o PLN, utilizando dos conhecimentos de linguística, nos permite aproveitar ao máximo o conteúdo do texto, oferecendo a possibilidade de correção de erros ortográficos, reconhecimento de sinônimos, remoção de ambiguidades. Porém, Witten et al. (2011) alerta de que a maioria das tarefas de PNL não são simples e requerem bastante suporte computacional.

Nesse sentido, Bird e Loper (2004) propuseram a biblioteca *Natural Language Process Toolkit* (NLTK) da linguagem Python para auxilio na tarefa de PLN, atualmente, na versão 3.3, o NLTK (2015), além de ter suporte para língua portuguesa, possui um conjunto de ferramentas para processamento de texto, classificação de texto, *tokenização*[2], remoção de *stop words*[3] e outros. Por isso, boa parte do trabalho pesado de PLN será realizado através do uso desse poderoso *framework*.

2.2.3 TEORIA DOS CLASSIFICADORES

Os classificadores utilizados nesse estudo foram escolhidos com base em seu uso para classificação em trabalhos semelhantes como o de Lima (2018), Cavalcante (2017) e Almeida et al. (2016). Os mesmos possuem bons resultados em trabalhos da área de mineração de texto. Nessa sessão, portanto, são observados os conceitos básicos dos classificadores definidos para essa pesquisa.

Naive Bayes é um classificador comum, estatístico, baseado no teorema de Bayes, demonstrado na equação (2.1):

$$P(c|f1 \dots fn) = \frac{P(c)P(f1 \dots fn|c)}{P(f1 \dots fn)} \quad (2.1)$$

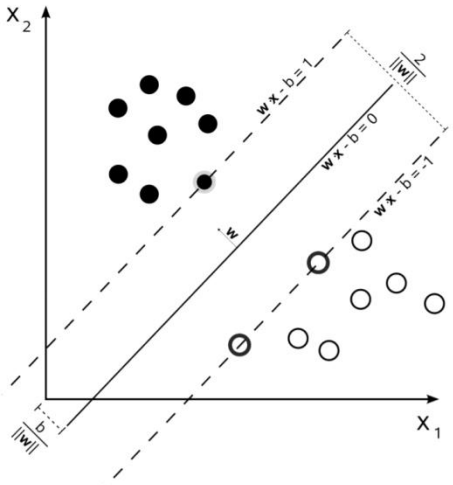
Sendo, c o atributo classe e $fn's$ os atributos levados em consideração para a classificação. Na classificação, a probabilidade $P(c|f1 \dots fn)$ é calculada para cada

[2] Processo de transformação de sentenças inteiras em *tokens* referentes a cada palavra da sentença
[3] Palavras sem nenhum valor semântico relevante.

classe c, assim sendo, o classificador identificará o elemento como pertencente a classe com maior probabilidade; A implementação escolhida foi a *MultinomialNB.*

Suporte Vector Machine (SVM) é um classificador também muito utilizado por ser efetivo em classificação de texto. O mesmo é baseado na teoria de aprendizagem estatística, desenvolvida por Vapnik (2000). Em um caso de duas classes, como é o deste trabalho, a ideia básica é a de que durante o treino do classificador, é gerado um hiperplano representado por um vetor $\vec{w}$, que não só separa uma classe da outra, mas que garante que a separação/margem seja o mais larga possível. O que corresponde a um problema de otimização restrito, ou seja, $c_j \in \{1, -1\}$ (onde, 1 corresponde a classe positiva e -1 negativa). A Figura 4 ilustra um exemplo de hiperplano com duas classes para classificação:

Figura 4. Exemplo de hiperplano do SVM



Fonte: Duarte (2013)

Sendo assim, a posição de uma entrada no teste, quão próxima ou distante do ponto definido pelo treino, indica ao classificador a que classe essa entrada pertence. A equação que define a teoria é a seguinte:

$$\vec{w} = \sum_j \alpha_j c_j \vec{d}_j \,,\ \alpha_j \geq 0, \qquad (2.2)$$

Onde, α_j's são obtidos resolvendo o problema de otimização de duas classes, conforme Pang et al. (2002), por fim $\vec{d}_j$ tal que $\alpha_j \geq 0$, são chamados de vetores de suporte, dessa forma o trabalho do classificador consiste em determinar de que lado do hiperplano de $\vec{w}$ a entrada de teste se encontra. A implementação escolhida para os testes foi a *SVC,* disponibilizada pelo *sci-kit*.

Ainda, segundo Pang et al. (2002) e Duarte (2013), a abordagem baseada em **Maximum Entropy** e Regressão Linear é um modelo de classificação semelhante ao Naive Bayes, porém baseado numa expressão exponencial:

$$P(c|f) = \frac{1}{N(fs)} exp(\sum_i \lambda_{i,c} F_{i,c}(f,c)) \quad (2.3)$$

Onde, $N(fs)$ se trata de uma função de normalização para as fs características; $\lambda_{i,c}$ é uma função de "peso", que muda durante o treinamento para uma classe c e $F_{i,c}(f,c)$ é uma função booleana que diz se aquelas fs características pertencem ou não a classe c. Como no Naive Bayes, a classe c, que tiver a maior probabilidade $P(c|f)$ é considerada a classe mais provável para f. E diferentemente do Naive Bayes, Maximum Entropy e/ou Regressão Linear não assumem que exista independência entre as classes. Dessa forma, definimos como implementação para testes o *SGDClassifier*.

2.3 FREQUÊNCIA LINEAR DAS PALAVRAS ($TF \times IDF$)

$TF \times IDF$ é um índice discriminante, oriundo de modelos de análise discriminante estatística baseada em conceitos Bayesianos, que, em linhas gerais, procura achar as palavras que mais discriminam o conjunto do documento analisado. Durante esse processo de aplicação do índice na coleção de dados, são atribuídos "pesos" aos termos/palavras, baseado na frequência desse termo na coleção de dados de texto.

Term frequency – TF é a medida da frequência do termo t_j no documento d_i, a ideia básica, segundo Aranha (2007), é a de que os termos que mais aparecem possuem maior relevância/peso do que aqueles que aparecem menos frequência no documento. Assim sendo, atribui-se a a_{ij} o valor $TF(t_j, d_i)$, conforme equação (2.4).

$$a_{i,j} = TF(t_j, d_i) \tag{2.4}$$

Vem ao caso observar que um termo frequente pode aparecer em toda a coleção de dados, o que não daria uma boa medida para de discriminação para a dar suporte a tarefa de classificação.

Nesse sentido, portanto, ainda segundo Aranha (2007), *Inverse document frequency* – IDF é uma medida que varia inversamente ao número de documentos que contém o termo/palavra t_j em um conjunto de documentos N. Assim sendo, essa medida, representada na equação (2.5) pode ser utilizada para dar um peso menor ao problema citado anteriormente.

$$IDF = \log \frac{N}{c} \tag{2.5}$$

Dessa forma, surge então a medida $TF \times IDF$, uma combinação de TF e IDF, equação (2.6).

$$a_{if} = TFIDF(t_j, d_i) = TF(t_j, d_i) \times \log \frac{N}{c} \tag{2.6}$$

2.3.1 MATRIZ DE CONFUSÃO

Matriz de confusão, representada no Quadro 1, como o próprio nome diz é uma representação da classificação feita, levando em consideração os acertos e a “confusão” feita pelo classificador.

Quadro 1. Matriz de confusão

Classe predita / Classe real	**Positivo**	**Negativo**
Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Negativo	Falso Negativo (FN)	Verdadeiros Negativos (VN)

Fonte: Elaboração própria

Assim, pode-se ter uma visão geral do número absoluto de acertos e erros que o classificador cometeu. Além disso, são esses números que ao serem equacionados fornecem uma porcentagem métrica do desempenho do algoritmo, conforme a seção seguinte aborda.

2.4 MÉTRICAS E ABORDAGENS PARA AVALIAÇÃO DOS CLASSIFICADORES

Diversas são as métricas que podem ser reconhecidas a partir da observação da matriz de confusão, utilizou-se: acurácia, precisão e revocação. Essas métricas são amplamente comumente utilizadas em trabalhos que avaliam desemprenho de classificadores, como é o caso dos trabalhos de Lima (2018), Cavalcante (2017) e Almeida et al. (2016), por exemplo.

A acurácia se trata de uma avaliação geral dos acertos, realizando uma relação com o total de instâncias do *dataset*.

$$Acurácia = \frac{Verdadeiros\ Positivos + Verdeiros\ Negativos}{Total} \quad (2.7)$$

Por sua vez, a precisão, se trata de uma avaliação entre os acertos em uma relação com a soma dos acertos e erros, respondendo ao questionamento '*daqueles que classifiquei como corretos, quantos efetivamente eram?'*, como podemos observar na equação 2.8.

$$Precisão = \frac{Veradeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Positivos} \quad (2.8)$$

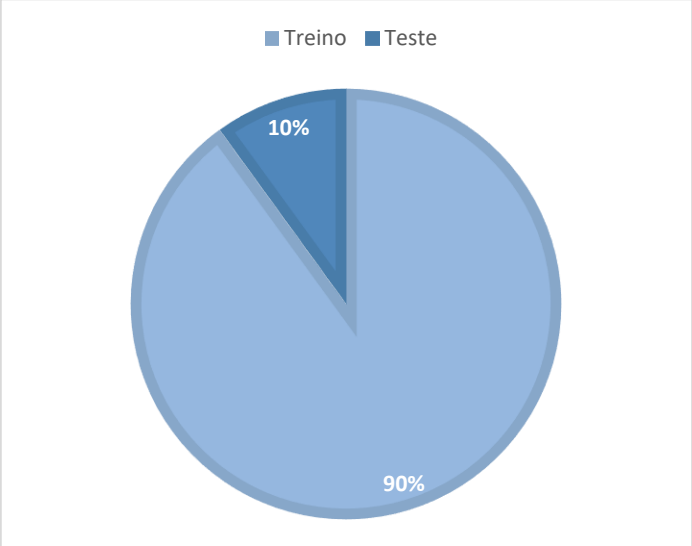
Por fim, a recall ou revocação se trata da medida que avalia a frequência com que o classificador classifica corretamente, respondendo ao questionamento *'quando a classe é X, quão frequente se classifica como X?'*. Observa-se isso, portanto, na equação 2.9.

$$Revocação = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos} \quad (2.9)$$

2.4.1 TREINO E TESTE

Treino e teste é uma abordagem utilizada para avaliações mais gerais e menos precisas sobre o desempenho de algoritmos na tarefa de classificação, onde o avaliador define uma porcentagem da base de dados para treino e outra para teste. Normalmente, a base é randomicamente organizada e dividida após essa organização, costuma-se utilizar entre 10% - 30% para teste e o restante para treino.

Figura 5. Distribuição da base para treino e teste



Fonte: Elaboração própria.

A Figura 5 ilustra a distribuição da base na proporção de 90% para treino e 10% para teste, assim, garantindo que os classificadores tenham maior quantidade de dados rotulados disponíveis para serem treinados e se tornarem o mais especialistas possível, dado o contexto desse trabalho.

2.4.2 VALIDAÇÃO CRUZADA

A validação cruzada é uma técnica que afere melhor o desempenho de um algoritmo, visto que, não avalia o desempenho do mesmo ao tentar classificar o mesmo conjunto teste, mas garantindo, portanto, que todo elemento da base faça parte do conjunto de testes e de treino.

Figura 6. Distribuição das iterações da validação cruzada

Fonte: Elaboração Própria

A Figura 6 demonstra a distribuição das iterações da validação cruzada com 5 iterações, dada uma organização aleatória da base de dados, a cada iteração têm-se um novo recorte para teste e treino. Assim, a cada iteração são aferidas métricas de avaliação do desempenho e, por fim, a média dessas métricas retorna uma avaliação mais precisa do desempenho do classificador.

2.5 BIBLIOTECAS E ETAPAS DE PRÉ-PROCESSAMENTO

Além da NLTK, já citada anteriormente, outras bibliotecas, todas escritas em Linguagem Python, deram suporte a realização desse trabalho para a etapa de pré-processamento e avaliação de desempenho dos algoritmos. Dentre as que foram utilizadas para a manipulação de dados, podemos destacar *Pandas*[4] e *Numpy*[5], ambas possuem funções semelhantes como conversão de formato de dados, recortes na base de dados (*select*), médias, resumos rápidos da coleção como número de linhas, colunas, médias, valor máximo, mínimo e outras funções.

Para a recuperação da base de dados dos tweets, utilizou-se a biblioteca *Tweepy*[6], como a própria descrição em sua página na internet diz “uma biblioteca fácil de usar para acessar a API do Twitter”, em tradução nossa. A *Tweepy* facilita o acesso à API do *Twitter*, que fornece arquivos *json* com as publicações referentes as requisições do coletor. Cada publicação vem em um arquivo único, que contém informações tanto do texto publicado como das interações na rede social como *retweets, favs,* dados de geolocalização e diversos outros.

Por isso, parte do pré-processamento já se inicia na limpeza desses arquivos, excluindo-se tudo que se considera desnecessário e transformando o que se deseja utilizar em *comma-separeted values* (CSV), valores separados por vírgula, em tradução literal, ou outro formato de arquivo de dados. No caso desse estudo em específico, fez-se uso da biblioteca através do script disponibilizado no trabalho de Cavalcante (2017).

Por outro lado, para o pré-processamento, somente da base de dados de *tweets* foi utilizada a biblioteca *Tweet-PreProcessor*[7] que deu suporte para as tarefas de remoção de caracteres específicos da rede social, como *hastags*, *emoticos*, *emojis*, marcações de usuários. Já para as tarefas de representação dos dados, avaliação

[4] https://pandas.pydata.org/
[5] http://www.numpy.org/
[6] http://www.tweepy.org/
[7] https://pypi.org/project/tweet-preprocessor/

dos algoritmos e uso das próprias implementações dos algoritmos, utilizou-se a biblioteca *Sci-Kit Learning*[8], amplamente utilizada para atividades desse tipo, a biblioteca em questão deu suporte para as atividades de avaliação e aplicação dos algoritmos definidos nas coleções.

Por último, fez-se uso da biblioteca *Matplotlib*[9] para geração dos gráficos de descrição das bases de dados e matriz de confusão na abordagem teste x treino. Essa biblioteca foi bem importante, pois pela praticidade, o *Jupyter Notebook*[10] foi utilizado. Se trata de uma aplicação web com suporte para dezenas de linguagens de programação, era possível visualizar os gráficos com a mudança feita, apenas recompilando o trecho que gerava o gráfico, sem precisar executar todo o código novamente.

2.5.1 TOKENIZAÇÃO

Uma etapa fundamental do pré-processamento é a *tokenização*, falando em linhas gerais, como dito por Witten et al. (2011), se trata da conversão de sentenças em palavras. Uma definição detalhada é dada por Aranha (2007) como o processo em que “o texto representado por uma sequência de caracteres é agrupado em um primeiro nível segundo fronteiras delimitadas por caracteres primitivos como espaço (“ ”), vírgula, ponto etc.” Nesse processo, um grupo de caracteres definido é chamado: *token.* Já um agrupamento de *tokens* é chamado: *tokenstream*.

[8] https://scikit-learn.org/stable/
[9] https://matplotlib.org/
[10] https://jupyter.org/

Figura 7. Tokenização

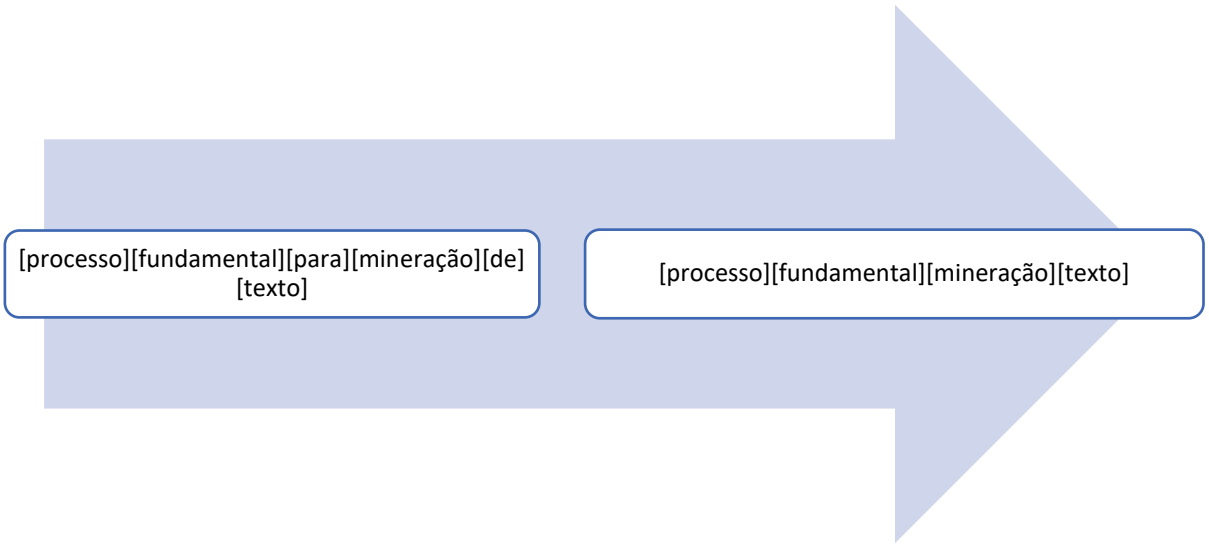
Fonte: Elaboração própria.

A Figura 7 ilustra o processo onde uma sentença é *tokenizada*, por assim se dizer, em que a cada ‘ ‘ (espaço) encontrado é identificado como o fim de um token e o próximo caractere encontrado, identificado como o início de outro.

2.5.2 REMOÇÃO DE *STOP-WORDS*

Witten et al. (2011) e Gupta; Lehal (2009) deixam claro em suas pesquisas que para evitar problema na tarefa de classificação, durante da fase de testes, deve-se, a priori, fazer a remoção de palavras que não carreguem valor semântico – *stop-words* –, exemplo disso são os artigos, que além de não possuírem valor semântico, se fazem muito presentes nos textos, o que pode induzir os classificadores ao erro. A Figura 8, exemplifica o processo de remoção de *stop-words*.

Figura 8. Remoção de stopwords



Fonte: Elaboração própria

Para essa etapa de pré-processamento a própria coleção padrão de *stopwords* para a língua portuguesa nativa do NLTK foi utilizada como parâmetro para a remoção das mesmas.

2.5.3 STEMMING

Stemming é o processo no qual uma palavra tem seu sufixo removido, restando apenas o radical, diminuindo assim, a quantidades de dados a serem processados no processo de classificação, porém mantendo-se a essência daquele termo em questão para o processo de “vetorização” (dotação de peso/importância) para o treinamento dos classificadores.

2.6 TRABALHOS RELACIONADOS

Cavalcante (2017) desenvolve um trabalho focado na criação de um *dataset* de rotulação automática de sentimentos baseada em emoticons para a língua portuguesa. Obtém bons resultados com os classificadores – *Support Vector Machines*, *Naive Bayes* e *Maximum Entropy* –, ainda mais se tratando de uma rede social – o Twitter – com tanto ruído nos textos, o que dificulta a tarefa. Por isso, regras específicas quanto a validade desses tweets para o *dataset* foram definidas, regras essas que são exploradas no capítulo do estudo de caso, na seção de descrição das coleções de dados.

Lima (2018) também avalia desempenho de algoritmos na tarefa de classificação de para coleções de rótulos polares, porém seu trabalho utiliza coleções de texto escrito em língua inglesa, sendo as duas coleções, uma sobre comentários de hotéis e outra de filmes, a versão original da coleção que de filmes que utilizou-se nesse trabalho. A pesquisa mostra que os classificadores utilizados *Support Vector Machines*, *Naive Bayes* e Regressão Linear – sendo que os dois primeiros também são utilizados também neste trabalho, mantém o desempenho similares para a coleção de dados de revisão de filmes, melhor definida no Capítulo 3.

Duarte (2013) realiza análise de desempenho de algoritmos para o domínio da rede social Twitter, porém, assim como no caso dessa pesquisa para língua portuguesa. Um diferencial importante é que a avaliação de desempenho de classificadores baseados em aprendizado de máquina se deu usando três atributos classes, sendo eles: negativo, positivo e neutro. Além disso, uma combinação de abordagens de análise de sentimentos foi utilizada para a extração de tópicos relevantes na rede social.

Por fim, esse trabalho se diferencia dos demais por ter como objeto de estudo coleções de dados que foram traduzidas automaticamente, além disso, por relacionar o desempenho dos classificadores em uma coleção de dados de domínio específico

com o desempenho desses algoritmos em uma coleção de dados ruidosa e de domínio não específico, que é o caso da coleção de dados textuais oriundos do Twitter.

3 ESTUDO DE CASO

Nessa seção é descrito em detalhes o estudo de caso de classificação da polaridade de opinião. Faz-se necessário relatar também alguns conceitos teóricos acerca das técnicas utilizadas durante essa fase da pesquisa, além de métodos de construção dos classificadores e uma breve contextualização sobre a coleção de dados obtida. E, claro, os testes e seus resultados.

3.1 DESCRIÇÃO DAS COLEÇÕES DE DADOS

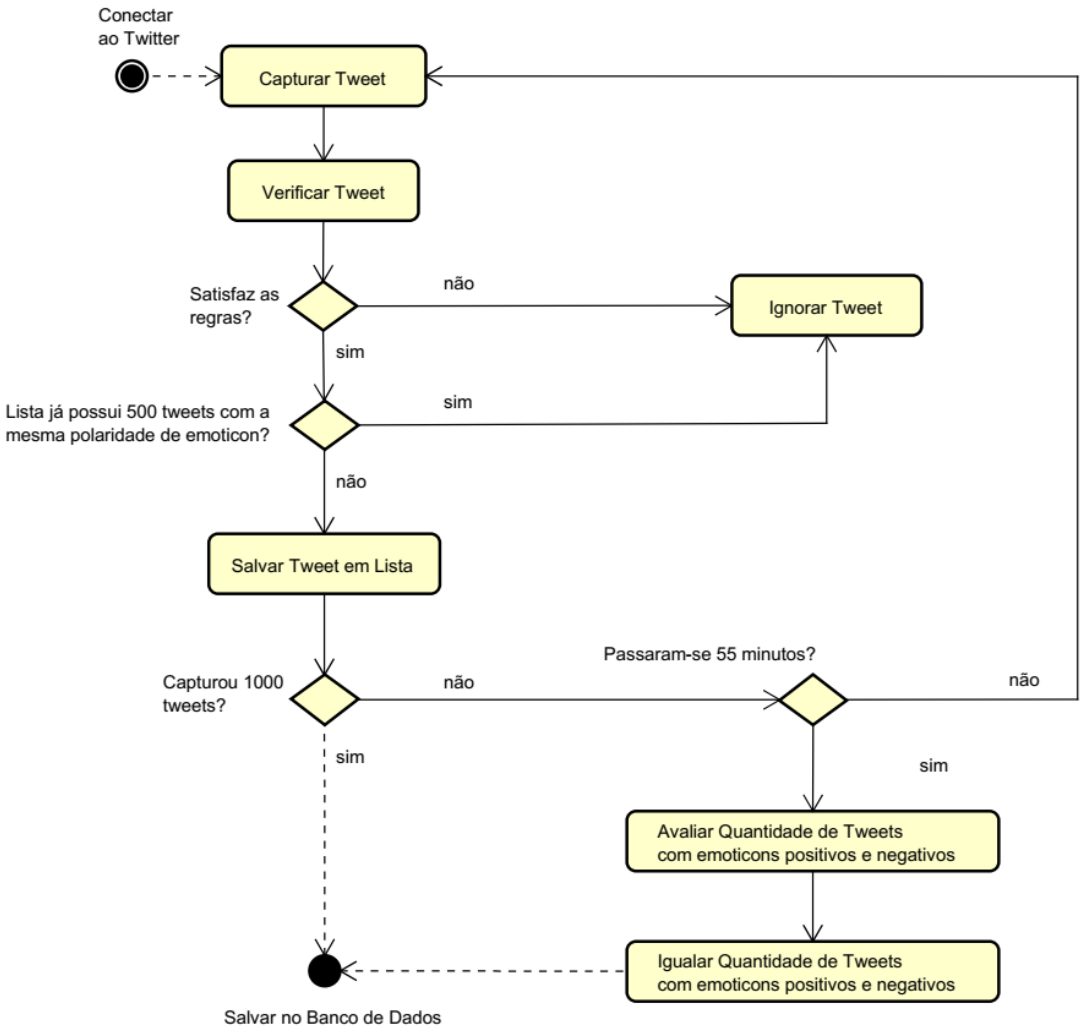
Nessa pesquisa, fez-se uso de 4 bases de dados para experimentações, sendo, 2 recortes, um balanceado e outro desbalanceado do *dataset* recuperados da rede social Twitter[11], já classificados e disponibilizados através do trabalho de Cavalcante (2017), e 2 recortes da base de dados de revisões de filmes traduzidas automaticamente do inglês para português[12].

Esses *dataset*s foram definidos por já terem sido objeto de estudo anterior, como no próprio trabalho que criou o *dataset* rotulado de tweets, como também o

[11] https://github.com/pauloemmilio/dataset
[12] https://www.kaggle.com/luisfredgs/imdb-ptbr

dataset de revisão de filmes que é utilizado no trabalho de Lima (2018) e Pang et al. (2002). O diferencial deste trabalho é utilizar esse *dataset* traduzido para a língua portuguesa.

Figura 9. O processo de coleta dos tweets



Fonte: Baseado em Cavalcante (2017)

A Figura 9 contém o fluxo do processo de coleta de tweets estabelecidos por Cavalcante (2017) regras para a classificação automática dos tweets estão descritas no Quadro 2.

Quadro 2. Regras para coleta de tweets

Regra	Exemplos	
	Aceito	**Recusado**
Deve conter ao menos um dos seguintes Emoticons: “:)”, “:-)”, “:(“, “:-(“	Estou feliz! :)	Estou feliz!
Não pode conter um Emoticons positivos e negativos ao mesmo tempo	Oi @ :)	:(Oi, @ :)
O idioma deve ser o português	Olá! :)	Hello! :)
Não pode ser composto apenas por Emoticons		:)
Não pode ser composto apenas por links acompanhados de Emoticon	Olha no www.google.com kkk :)	www.google.com :)
Não pode ser composto apenas por nomes de usuários acompanhados de Emoticons	Oi @JardelCunha :)	@JardelCunha
Não podem possuir o mesmo texto. Implicando em recusar retweets[13] ou tweets com mesmos IDs.	Bom dia Brasil, Boa tarde Itália :)	Bom dia Brasil, Boa tarde Itália :)

Fonte: Elaboração própria.

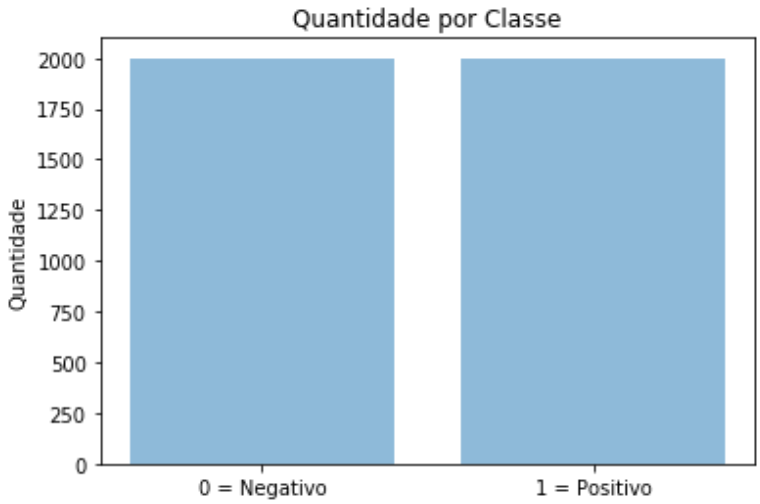
Para o *dataset*, apenas os dados referentes ao corpo do *tweet*, foram levados em consideração, ignorando-se demais dados do arquivo *json*. Foi possível recuperar 25983 rotuladas como positivos e 27287 rotulados como negativos. Um número menor do que os 75828 do *dataset* original, uma vez que as publicações estão sujeitas a serem excluídas da base de dados da rede social por fatores como exclusão das próprias contas de usuário, dos *tweets* e mudanças nas escolhas de privacidade dos usuários autores dos tweets.

Como objeto de estudo, por capacidade limitada de processamento da máquina utilizada para o experimento, realizou-se um recorte com aproximadamente 4000 linhas da base dados, sendo 50% delas rotulas como positiva e 50% negativas. Conforme o gráfico da Figura 8. Para efeito de organização dos resultados exibidos posteriormente, chamaremos essa base de dados de *Tweets-Balanceada*.

A seguir, têm-se as ilustrações gráficas das distribuições das classes em cada base de dados do escopo das experimentações, são 4, oriundas de 2 domínios, sendo que cada domínio possui uma base balanceada e outra desbalanceada.

[13] Funcionalidade da rede social que permite repostar um tweet de outro usuário.

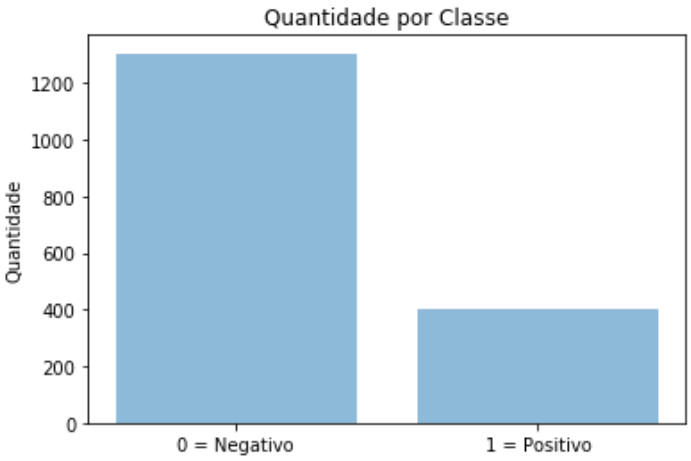
Figura 10. Base de dados de tweets balanceada



Fonte: Elaboração própria

A base referência da Figura 10 possui como rótulo para negativo ‘0’ e ‘1’ para positivo. A base em questão possuía, originalmente, muito ruídos contidos no texto, oriundos da própria rede social e também do comportamento específico que usuário adotam no seu uso. Essa especificidade acarreta um maior esforço no pré-processamento para esses dois primeiros recortes de coleções de dados.

Figura 11. Base de dados de tweets desbalanceada

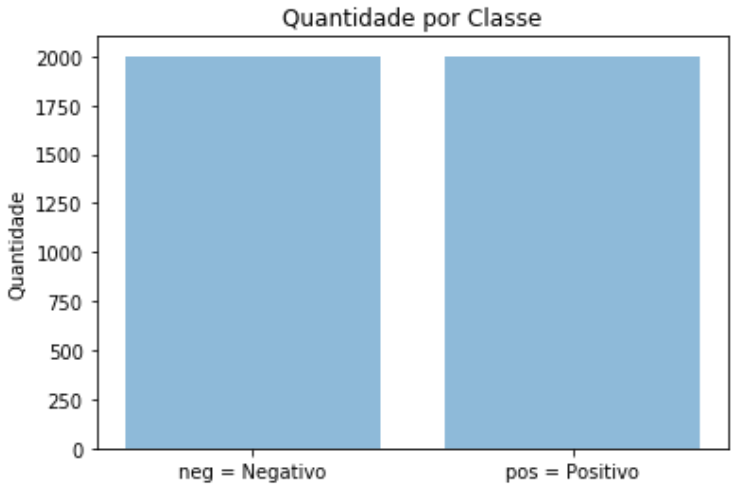


Fonte: Elaboração própria

A Figura 11 ilustra a distribuição do *dataset* de tweets desbalanceado, que por se tratar de um recorte do *dataset* balanceado, possui tamanho inferior aos demais *datasets*.

Já a Figura 12 mostra a distribuição por classe negativa e positiva, com rótulos ‘neg’ para sentenças negativas e ‘pos’ para sentenças positivas, de uma base de dados de revisões de filmes feitas em comentários do IMDb[14]. A distribuição se deu de igual forma. E igualmente, por organização, essa base de dados foi nominada como *IMDb-Balanceada.*

Figura 12. Base de dados de revisões de filmes balanceada

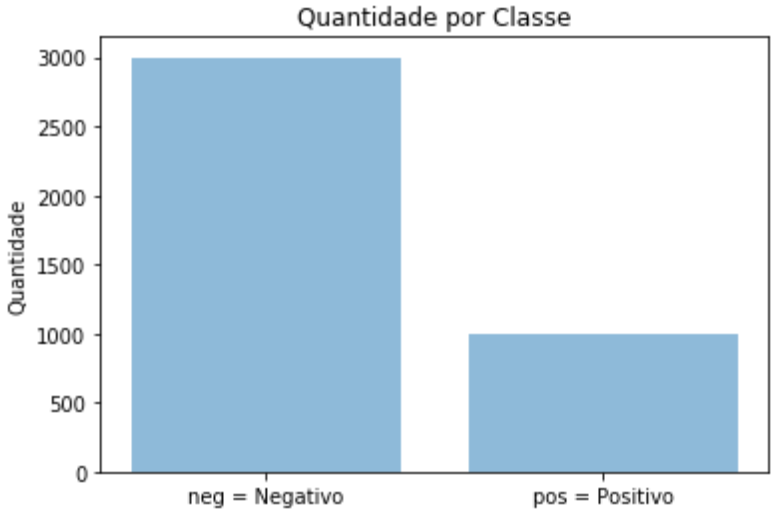


Fonte: Elaboração própria

Assim, segue-se com *IMDb-desbalanceada* e suas distribuições por classe são ilustradas na Figura 13.

[14] https://www.imdb.com/

Figura 13. Base de dados de revisões de filmes desbalanceada



Fonte: Elaboração própria

3.2 PRÉ-PROCESSAMENTO

O pré-processamento é necessário para transformar o problema de mineração de opinião em um problema de classificação, dessa forma, os dados desestruturados precisam ser processados através das técnicas citadas no Cap. 2. Assim, conseguiu-se criar uma coleção que pode ser tratada como um problema de classificação. Utilizando recursos das bibliotecas citadas anteriormente realizaram-se diversos processos de limpeza das bases de dados, dentre os processos destacam-se:

- Remoção de pontuação;
- Remoção de emoticons e emojis;
- Remoção de hashtags (tags de indexação utilizada no twitter);
- Remoção de URL’s;

- Remoção de marcações de usuários;

Além dessas, outras atividades de pré-processamento foram realizadas e são descritas mais afundo nas subseções que seguem.

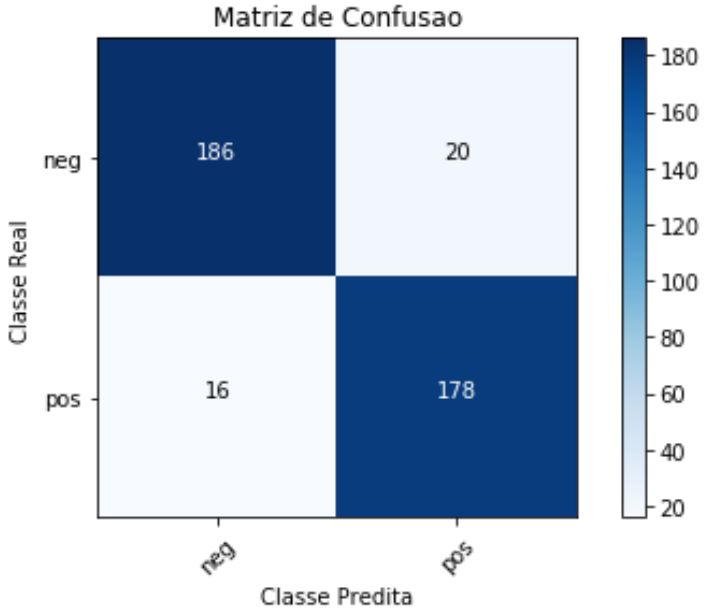
3.3 RESULTADOS

Nessa seção apresentam-se os resultados das experimentações, assim como, as discussões referentes aos mesmos. $TF \times IDF$ foi o índice discriminante utilizado durante todas as experimentações que estão descritas nas duas abordagens descritas nos resultados.

3.3.1 TREINO E TESTE

Durante a aplicação da abordagem de treino o teste, foi utilizada a proporção de 90% para treino e 10% para teste. As figuras que seguem exibem a matriz de confusão dessa abordagem, enumerando os acertos e erros dos algoritmos, além disso, é possível observar um quadro com o resumo geral das métricas para cada base testada.

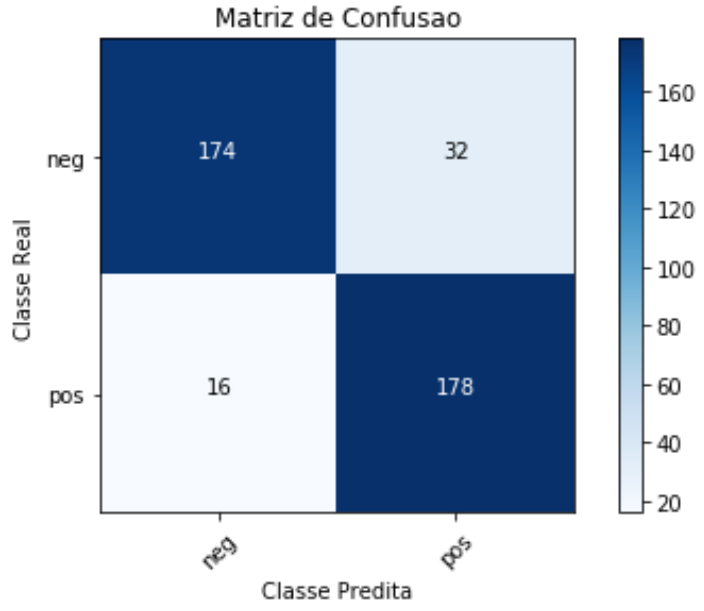
Figura 14. *Naive Bayes* para *IMDb-Balanceada*



Fonte: Elaboração própria

A Figura 14 apresenta a matriz de confusão para o classificador Naive Bayes na base *IMDb-Balanceada.* Já a Figura 15, para o algoritmo *SVM*.

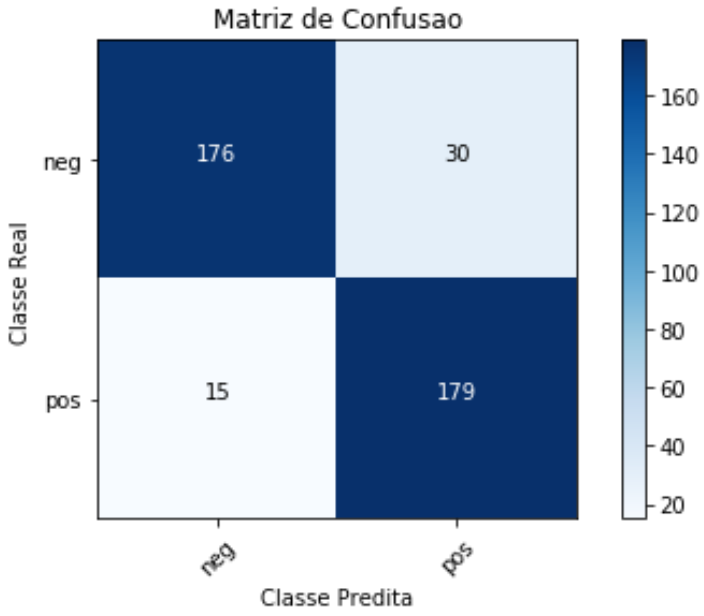
Figura 15. *SVM* para *IMDb-Balanceada*



Fonte: Elaboração prórpia

Na Figura 16 pode-se observar a matriz de confusão para o algoritmo *Maximum Entropy* para a coleção de revisões de filmes desbalanceada.

Figura 16. *Maximum Entropy* para *IMDb-Balanceada*



Fonte: Elaboração própria

O Quadro 3 traz um resumo das métricas de avaliação dos 3 algoritmos durantes os testes com a base de dados IMDb-Balanceada.

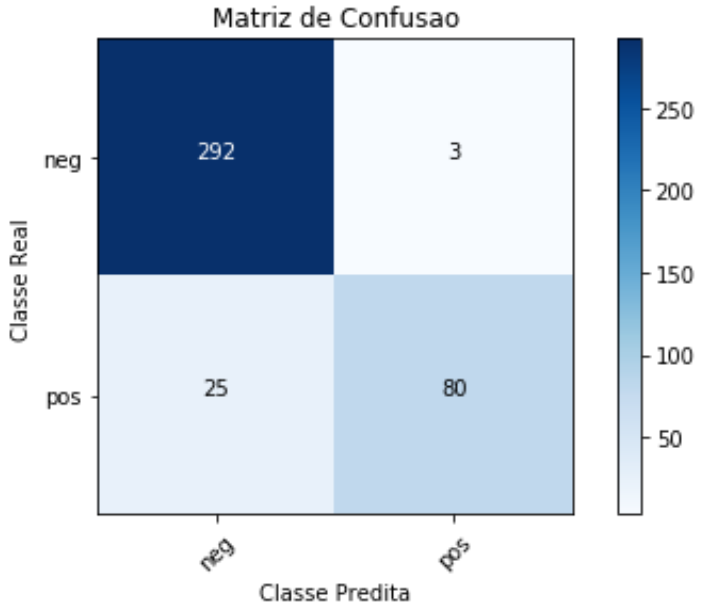
Quadro 3. Resumo das méticas para *IMDb-Balanceada*

ALGORITMO	**MÉTRICAS**				
	Acurácia	Precisão		Revocação	
		neg	pos	neg	pos
Naive Bayes	0.91	0.92	0.9	0.9	0.92
SVM	0.88	0.92	0.85	0.84	0.92
Maximum Entropy	0.88	0.92	0.86	0.85	0.92

Fonte: Elaboração prórpia

Podemos observar que *Naive Bayes* obteve um desempenho consistente e significativo, conseguindo classificar bem para ambas as classes, apesar do desempenho semelhante dos demais classificadores, ainda assim, se saiu melhor. A Figura 17 traz a matriz de confusão para o *Naive Bayes* aplicado a base *IMDb-Desbalanceada.*

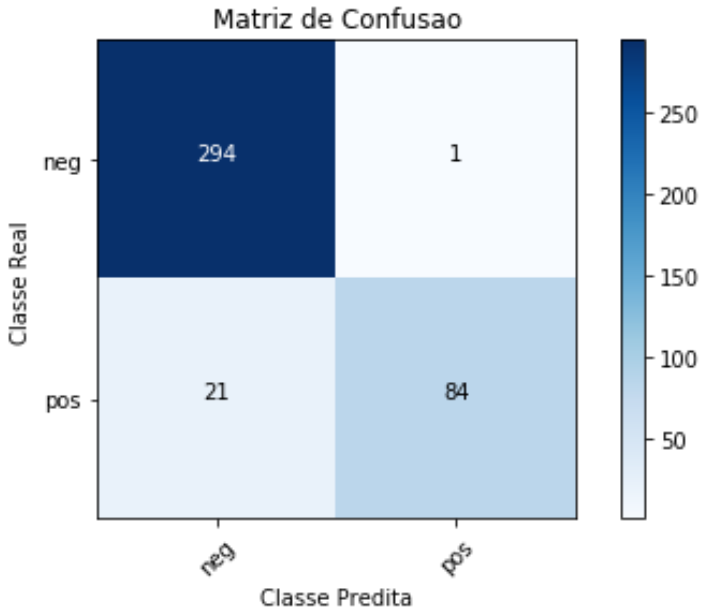
Figura 17. *Naive Bayes* para *IMDb-Desbalanceada*



Fonte: Elabooração prórpria

Em seguida, temos a Figura 18 com os acertos e erros do *SVM* junto a mesma base.

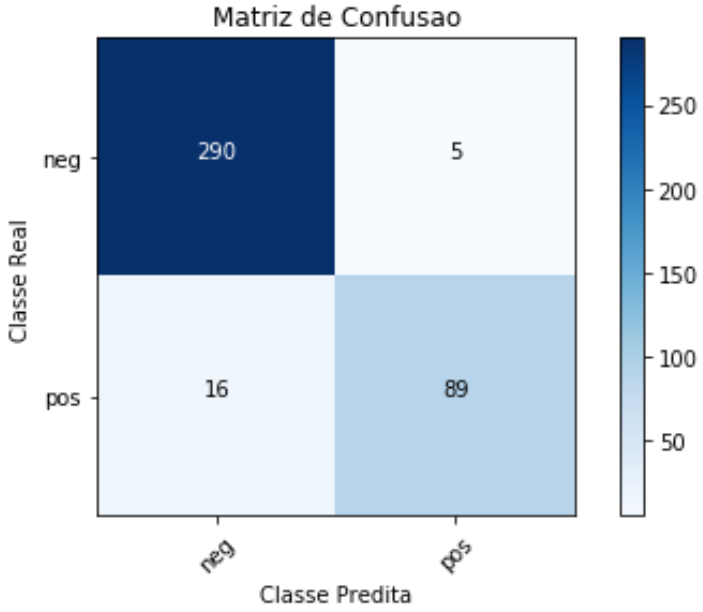
Figura 18. SVM para IMDb-Desbalanceada



Fonte: Elaboração própria

Logo abaixo a Figura 19, matriz de confusão do resultado de classificação do *Maximum Entropy,* também para a base desbalanceada de revisão de filmes.

Figura 19. *Maximum Entropy* para *IMDb-Desbalanceada*



Fonte: Elaboração prórpria

A seguir, o Quadro 4 com um resumo das métricas para o *dataset* desbalanceado das revisões sobre os filmes.

Quadro 4. Resumo das métricas para IMDb-Desbalanceada

CLASSIFICADOR	**MÉTRICAS**				
	Acurácia	Precisão		Revocação	
		neg	pos	Neg	pos
Naive Bayes	0.93	0.92	0.96	0.99	0.76
SVM	0.94	0.93	0.99	1	0.8
Maximum Entropy	0.94	0.94	0.96	0.99	0.83

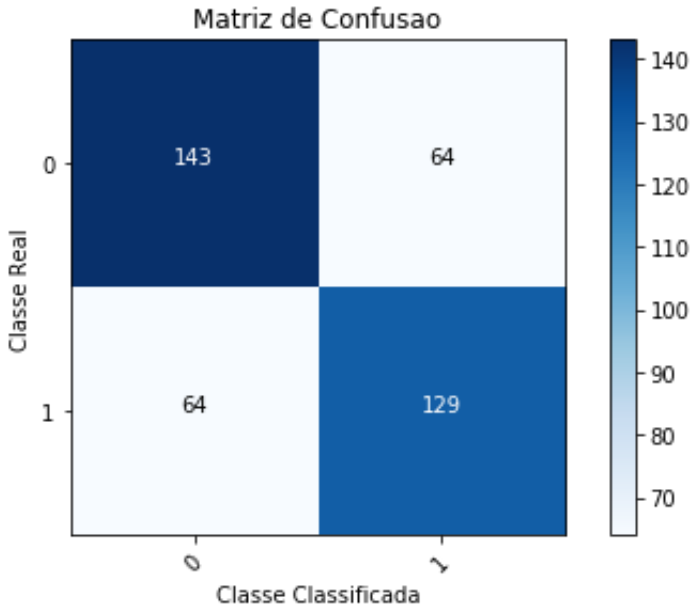
Fonte: Elaboração própria

Pode-se observar que o *Naive Bayes* perde em desempenho para o *SVM* quando se trata de *dataset* com quantificação de rótulos/classe desbalanceadas,

assim sendo, o último obteve melhor desempenho para esse cenário. Além disso, o *Maximum Entropy* obteve resultados superiores em precisão para classe negativa e em revocação para classe positiva, como também resultados semelhantes em acurácia e nas demais classes e métricas.

Dando continuidade aos resultados para as demais coleções, temos em seguida a matriz de confusão do *Naive Bayes* junto ao recorte balanceados de tweets na Figura 20.

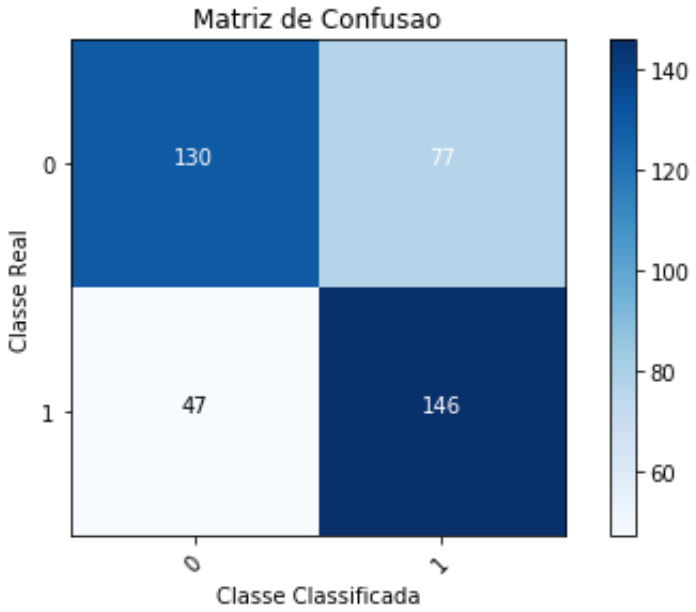
Figura 20. *Naive Bayes* para *Tweets-Balanceada*



Fonte: Elaboração própria

E, logo após, a Figura 21, traz os resultados de *SVM* para o recorte balanceado de postagens do *microblog*.

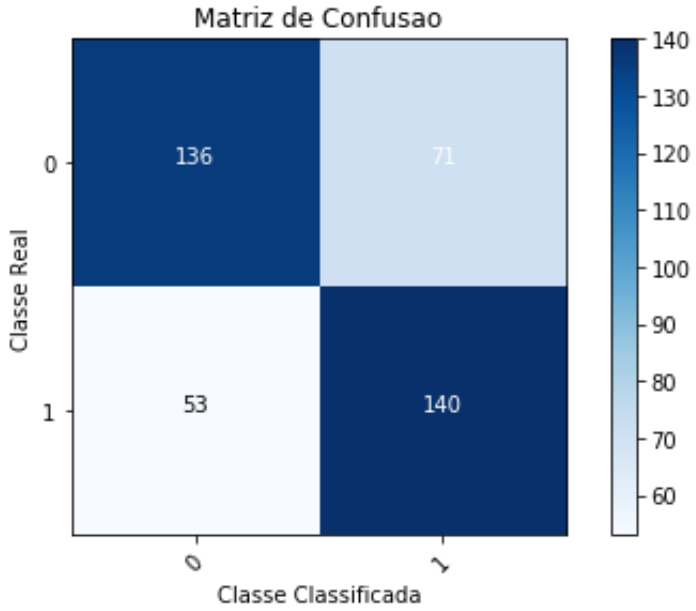
Figura 21. *SVM* para *Tweets-Balanceada*



Fonte: Elaboração própria

Por fim, a Figura 22 ilustra a matriz de confusão dos acertos e erros do *Maximum Entropy* para a coleção balanceada de tweets.

Figura 22. *Maximum Entropy* para *Tweets-Balanceada*



Fonte: Elaboração própria

Abaixo, o Quadro 5 traz o resumo das métricas que retratam o desempenho dos classificadores na coleção citada anteriormente.

Quadro 5. Resumo das métricas para Tweets-Balanceada

CLASSIFICADOR	**MÉTRICAS**				
	Acurácia	Precisão		Revocação	
		neg	pos	neg	pos
Naive Bayes	0.68	0.69	0.67	0.69	0.67
SVM	0.69	0.73	0.65	0.63	0.76
Maximum Entropy	0.69	0.72	0.66	0.66	0.73

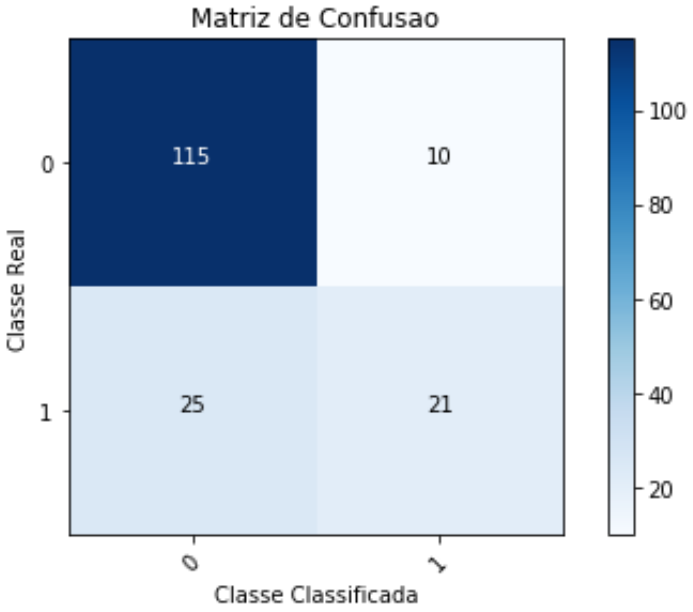
Fonte: Elaboração pória

Comparado aos desempenhos dos mesmos classificadores com os resultados da coleção de dados de revisão de filmes, o resultado está abaixo dos obtidos com os dados sobre filmes, isso se dá devido à falta de especificidade sobre o que se tratam os textos da rede social, conforme explicado por Liu (2012). Ainda assim, resultados de acurácia próximos a 70% e de precisão e revocação que passam disso, comprovam a eficiência dos classificadores. Ainda mais por se tratarem de dados textuais que não retratam sentimentos a cerca de uma entidade ou evento específico, mas estão relacionados a expressões textuais que contém sentimento acerca dos mais diversos temas e aspectos.

Assim sendo, pode-se observar no quadro acima que *Maximum Entropy* e SVM obtiveram resultados equivalentes e *Naive Bayes* o menor desempenho, apesar da proximidade percentual do desempenho do *SVM*.

A seguir, temos matriz de confusão para o *Naive Bayes* com a coleção desbalanceada de postagens do microblog na Figura 23.

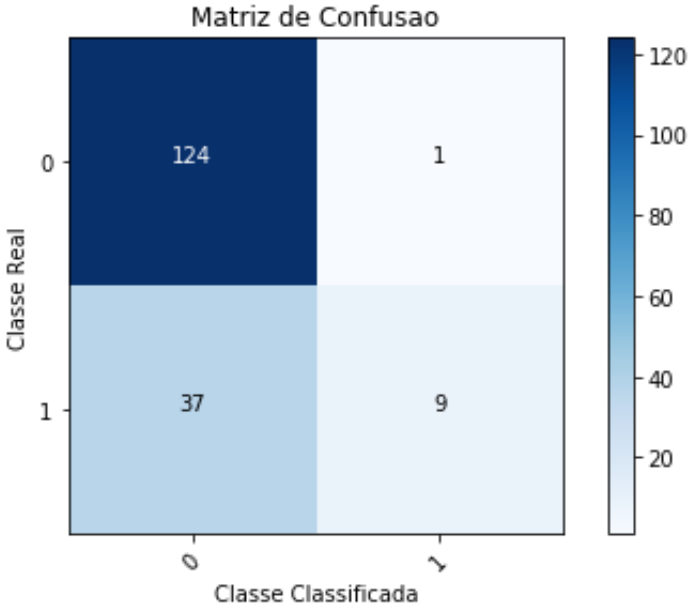
Figura 23. *Naive Bayes* para *Tweets-Desbalanceada*



Fonte: Elaboração própria

E logo mais abaixo, na Figura 24, a matriz de confusão de *SVM* para o teste em *Tweets-Desbalanceada.*

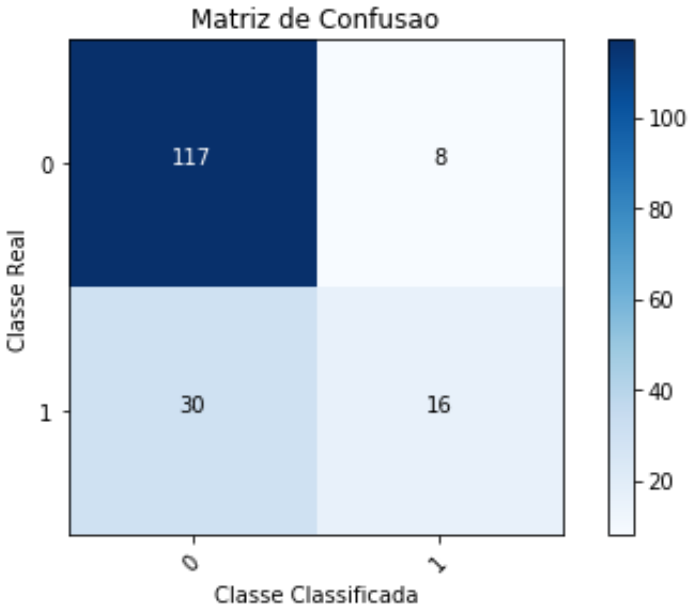
Figura 24. *SVM* para *Tweets-Desbalanceada*



Fonte: Elaboração própria

Por fim, na Figura 25, a matriz de confusão para *Maximum Entropy* no teste com o recorte do *dataset* de tweets desbalanceado.

Figura 25. *Maximum Entropy* para *Tweets-Desbalanceada*



Fonte: Elaboração Própria

O Quadro 6 reúne as métricas para os 3 classificadores nos testes com *Tweets-Desbalanceada*.

Quadro 6. Resumo das métricas para *Tweets-Desbalanceada*

CLASSIFICADOR	**MÉTRICAS**				
	Acurácia	Precisão		Revocação	
		neg	pos	neg	pos
Naive Bayes	0.79	0.82	0.68	0.92	0.46
SVM	0.77	0.77	0.09	0.99	0.2
Maximum Entropy	0.77	0.94	0.35	0.8	0.67

Fonte: Elaboração prórpia

É possível notar uma queda significativa no desempenho dos classificadores em relação a coleção balanceada, mesmo com a alta da acurácia, é evidente a baixa taxa de precisão para classe negativa, ao olhar para as matrizes de confusão, vê-se que todos mais erram do que acertaram quando se tratava do rótulo com menos instâncias em treino. Além disso, a frequência com que se classificava corretamente é bem baixa, vide a métrica revocação, o pior desempenho foi o do *SVM*. Aquele que teve a menor perda de desempenho com relação a coleção balanceada foi o *Naive Bayes*.

Nesse sentido também, é importante observar que nos testes com as coleções de revisões de filmes o desempenho teve menor impacto, com relação ao balanceamento, do que com as coleções de tweets, o que corrobora com o que diz Liu (2012) sobre a o desempenho dos algoritmos serem melhores em domínios mais específicos, ao se tratar da análise de sentimentos.

3.3.2 VALIDAÇÃO CRUZADA

A validação cruzada, foi feita com a mesma proporção de treino e teste. Utilizou-se a abordagem de 5 iterações, assim, foram feitas as médias das métricas e compiladas nos quadros que seguem.

O Quadro 7 apresenta a compilação dos resultados da validação cruzada. A validação cruzada traz uma avaliação mais precisa e um refinamento quanto às métricas. Nesse sentido, podemos observar que não ouve discrepância alguma do método de treino e teste, porém algumas observações se fazem importante de serem realizadas:

- *Maximum Entropy* é o classificador que obteve melhor desempenho quando leva-se em consideração o balanceamento ou não das coleções;
- *Naive Bayes* e *SVM,* vide revocação, perdem desempenho significativo quando observa-se o balanceamento das coleções, sendo o *SVM* o que

possui maior perda de desempenho quando se tratam de coleções desbalanceadas.

Quadro 7. Resultados para validação cruzada

Classificador	Métrica / Base	Acurácia	Precisão	Revocação
Naive Bayes	*IMDb-Balanceada*	0.72	0.73	0.72
	IMDb-Desbalanceada	0.78	0.75	0.60
	Tweets-Balanceada	0.68	0.68	0.68
	Tweets-Desbalanceada	0.78	0.69	0.65
SVM	*IMDb-Balanceada*	0.76	0.76	0.76
	IMDb-Desbalanceada	0.80	0.80	0.63
	Tweets-Balanceada	0.68	0.68	0.68
	Tweets-Desbalanceada	0.79	0.75	0.62
Maximum Entropy	*IMDb-Balanceada*	0.77	0.77	0.77
	IMDb-Desbalanceada	0.81	0.81	0.65
	Tweets-Balanceada	0.68	0.68	0.68
	Tweets-Desbalanceada	0.78	0.72	0.62

Fonte: Elaboração própria

4 CONSIDERAÇÕES FINAIS E RECOMENDAÇÕES

O presente trabalho de pesquisa realizou análise de desempenho de algoritmos classificadores na tarefa de classificação da polaridade num contexto de análise de sentimentos. As sessões seguintes versam sobre as considerações finais desta pesquisa e recomendações para trabalhos futuros.

4.1 CONSIDERAÇÕES FINAIS

Nesta pesquisa foram realizadas várias tarefas num contexto de mineração de opinião para que fosse possível avaliar o desempenho de algoritmos na classificação da polaridade de sentimentos em tweets e revisões de filmes. Passando primeiramente pela revisão de literatura para que então fosse possível realizar-se: a escolha das coleções, recortes, alteração de balanceamentos, atividades de pré-processamento, seleção dos classificadores utilizados, definição de métricas e o experimento em si. Após isso foram gerados ilustrações e quadros com o resumo dos resultados para que os mesmos pudessem ser discutidos.

Alcançando-se, então, a resposta para a questão-tema dessa pesquisa, visto que, apesar das quedas de desempenho com coleções desbalanceadas, os três algoritmos mantiveram desempenho consistente, mas fica clara a diferença de desempenho dos mesmos quando os resultados são relacionados. Dessa forma, deve-se o devido destaque ao *Maximum Entropy* e *SVM*, frente ao *Naive Bayes*. Além disso:

- Observou-se melhores resultados na frequência de classificações corretas, vide revocação, nas coleções balanceadas;
- Como também, melhores resultados foram obtidos nas coleções de domínio específico, caso este, das duas coleções sobre revisões de filmes;
- Na validação cruzada *SVM e Maximum Entropy* obtiveram resultados bem próximos ou superiores a 70%, como:
    - *Maximum Entropy* 81% de precisão para *IMDb-desbalanceada* e *SVM* 80% para a mesma métrica e base;
    - Enquanto isso, *Naive Bayes* perde em 5 pontos percentuais no *dataset* citado anteriormente, além de não obter pontuação na casa dos 80% em nenhum cenário para precisão e revocação.

Dessa forma, levando em consideração o contexto desse trabalho, vê-se que a avaliação de desempenho de classificadores na tarefa de mineração de opinião/análise de sentimento é de grande importância para que se possa escolher o melhor modelo para uma aplicação em contextos de negócio inteligente, afinal, melhores classificações resultam em melhores retornos às entidades que utilizam dessa ferramenta. Assim sendo, avaliar o desempenho de classificadores torna-se uma etapa importante a ser desenvolvida previamente, dando maior segurança aos investidores nessa tecnologia.

4.2 RECOMENDAÇÕES

Ao fim desta pesquisa, do ponto de vista da ciência, percebeu-se que outros aspectos dessa área de pesquisa podem ser explorados, tais como:

1. Avaliar desempenho de classificação de bases com mais de dois rótulos de sentimentos;
2. Aplicação de métodos léxicos para remoção de possíveis ruídos em bases classificadas pela abordagem de emoticons;
3. Avaliar o impacto da introdução de ruídos;
4. Criar e avaliar *dataset* automático utilizando emojis de teclados de smartphones mais sofisticados e que representam outros sentimentos como: nojo e medo;
5. Criar *dataset* rotulado sobre sarcasmo e ironia para a língua portuguesa.

REFERÊNCIAS

ALMEIDA, T. G.; SOUZA, B. A.; MENEZES, A. F. A.; FIGUEREIDO, C. M. S; NAKAMURA, E. F. **Sentiment Analysis of Portuguese Comments from Foursquare**, 2016. Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web, p. 355-358. Disponível em: <https://dl.acm.org/citation.cfm?doid=2976796.2988180>. Acesso em: 22/02/2019.

ARANHA, C. N. **Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional**, 2007. Pontifica Universidade Católica do Rio de Janeiro. Disponível em: <http://eds.a.ebscohost.com/eds/detail/detail?vid=1&sid=1963d31b-9887-4b50-ad3f-5d3bc06dc934%40sessionmgr4008&bdata=Jmxhbmc9cHQtYnImc2l0ZT1lZHMtbGl2ZSZzY29wZT1zaXRl#AN=edsndl.oai.union.ndltd.org.IBICT.oai.agregador.ibict.br.BDTD.oai.bdtd.ibict.br.PUC.RIO.o>. Acesso em: 24/10/2018.

ARANHA, C.; PASSOS, E.; PASSOS, E. **A Tecnologia de Mineração de Textos.** Revista Eletrônica de Sistemas de Informação, v. 5, n. 2, 2006. Disponível em: <http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171>. Acesso em: 24/10/2018.

BIRD, S.; LOPER, E. **NLTK**. Proceedings of the ACL 2004 on Interactive poster and demonstration sessions -. Anais... . p.31–es, 2004. Morristown, NJ, USA: Association for Computational Linguistics. Disponível em: <http://portal.acm.org/citation.cfm?doid=1219044.1219075>. Acesso em: 25/5/2018.

CAVALCANTE, P. E. C. **Um dataset para análise de sentimentos na língua portuguesa.** , 2017. Universidade Federal da Paraíba. Disponível em: <https://repositorio.ufpb.br/jspui/handle/123456789/3237>. Acesso em: 5/12/2018.

CHEONG, M.; LEE, V. C. S. **A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to**

terrorism events via Twitter. Information Systems Frontiers, v. 13, n. 1, p. 45–59, 2011. Springer US. Disponível em: <http://link.springer.com/10.1007/s10796-010-9273-x>. Acesso em: 24/10/2018.

DUARTE, E. S. **Sentiment analysis on twitter for the portuguese language**. , 2013. Faculdade de Ciências e Tecnologia. Disponível em: <https://run.unl.pt/handle/10362/11338>. Acesso em: 5/2/2019.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. 6ª ed. São Paulo: Pearson Addison Wesley, 2011.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **The KDD process for extracting useful knowledge from volumes of data**. Communications of the ACM, v. 39, n. 11, p. 27–34, 1996. ACM. Disponível em: <http://portal.acm.org/citation.cfm?doid=240455.240464>. Acesso em: 23/8/2017.

GIL, ANTONIO C. G. **Como elaborar projetos de pesquisa.** Atlas, 2010.

GUPTA, V.; LEHAL, G. S. **A Survey of Text Mining Techniques and Applications - Volume 1**, No. 1, August 2009 - JETWI. Journal of Emerging Technologies in Web Intelligence, ago. 2009. Disponível em: <http://www.jetwi.us/index.php?m=content&c=index&a=show&catid=165&id=969>. Acesso em: 24/10/2018.

LIMA, V. H. DA S. **Análise de algoritmos supervisionados na tarefa de classificação da polaridade de revisões**, 2018. Universidade Federal do Acre.

LIU, B. **Sentiment Analysis and Opinion Mining**. Morgan & Claypool, 2012.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams Engineering Journal**, v. 5, n. 4, p. 1093–1113, 2014. Elsevier. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2090447914000550>. Acesso em: 1/6/2018.

NLTK. Natural Language Toolkit. **NLTK**, 2015. Disponível em: <http://www.nltk.org/>. Acesso em: 24/10/2018.

PANG, B. e LEE, L. Opinion Mining and Sentiment Analysis. **Foundations and Trends® in Information Retrieval**, v. 2, n. 1–2, p. 1–135, 2008. Now Publishers Inc. Disponível em: <http://www.nowpublishers.com/article/Details/INR-011>. Acesso em: 24/5/2018.

PANG, B. e LEE, L. e VAITHYANATHAN, S. **Thumbs up?: sentiment classification using machine learning techniques**. Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02. Anais... . v. 10, p.79–86, 2002. Morristown, NJ, USA: Association for Computational Linguistics. Disponível em: <http://portal.acm.org/citation.cfm?doid=1118693.1118704>. Acesso em: 11/5/2018.

SILBERSCHATZ, A., SUNDARSHAN, S. e KORTH, H. F. **Sistema de Banco de**

Dados. 6º ed. Rio de Janeiro, 2012.

TSYTSARAU, M. e PALPANAS, T. **Survey on mining subjective data on the web.** Data Mining and Knowledge Discovery, v. 24, n. 3, p. 478–514, 2012. Springer US. Disponível em: <http://link.springer.com/10.1007/s10618-011-0238-6>. Acesso em: 3/10/2018.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. New York, NY: Springer New York, 2000.

WAZLAWICK, R. S. **Metodologia de pesquisa para ciência da computação**. 2º ed. Rio de Janeiro: Elsevier Editora Ltda., 2014.

WITTEN, I. H., IAN H., FRANK, E., HALL, M. A., MARK A. e PAL, C. J. **Data mining : practical machine learning tools and techniques**. 3ª ed. Burlington: Elsevier/Morgan Kaufmann, 2011.