

Edeleon Marcelo Nunes Brito

**Mineração de Textos: Detecção automática de
sentimentos em comentários nas mídias
sociais**

Belo Horizonte-MG

2017

Edeleon Marcelo Nunes Brito

Mineração de Textos: Detecção automática de sentimentos em comentários nas mídias sociais

Dissertação apresentada ao Programa de Mestrado em Sistemas de Informação e Gestão do Conhecimento da Universidade Fundação Mineira de Educação e Cultura — FUMEC, como requisito parcial para a obtenção do título de Mestre em Sistemas de Informação e Gestão do Conhecimento.

Área de concentração: Gestão de Sistemas de Informação e do Conhecimento.
Linha de Pesquisa: Recuperação da Informação e Sistemas de Informação

Universidade FUMEC

Faculdade de Ciências Empresariais

Programa de Pós-Graduação *Stricto Sensu* — Mestrado em Sistemas de Informação e Gestão do Conhecimento

Orientador: Prof. Dr. Luiz Cláudio Gomes Maia

Belo Horizonte-MG

2017



Universidade FUMEC
Faculdade de Ciências Empresariais
Curso de Mestrado em Sistemas de Informação e Gestão do
Conhecimento

Dissertação intitulada “**Mineração de Textos: Detecção automática de sentimentos em comentários nas mídias sociais**”, de autoria do mestrando **Edeleon Marcelo Nunes Brito**, aprovada pela banca examinadora constituída pelos seguintes professores:

Prof. Dr. Luiz Cláudio Gomes Maia — Universidade FUMEC
(Orientador)

Prof. Dr. Rodrigo Fonseca e Rodrigues — Universidade FUMEC

Prof. Dr. Fernando Hadad Zaidan — IETEC

Prof. Me. Eduardo Cardoso Melo — IFMG
(Membro *ad hoc*)

Prof. Dr. Fernando Silva Parreiras
Coordenador do Curso de Mestrado em Sistemas de Informação e Gestão do
Conhecimento
Universidade FUMEC

Belo Horizonte-MG, 13 de Fevereiro de 2017

Agradecimentos

A Deus;

Ao meu orientador, professor, Dr. Prof. Luiz Cláudio Gomes Maia, pela dedicação e ajuda no desenvolvimento deste trabalho;

aos meus pais, pelo apoio e compreensão;

aos meus irmãos, pelo incentivo e companheirismo;

Aos companheiros que estiveram juntos durante toda esta caminhada, em especial meu amigo Erick que por muitas vezes cedeu sua casa para que eu pudesse dormir;

Aos professores do Mestrado em Sistemas de Informação e Gestão do Conhecimento que me possibilitaram chegar até aqui;

A todos, muito obrigado.

“Se você se empenhar o suficiente, poderá fazer qualquer história resultar” (Saul Goodman, em Breaking Bad”)

Resumo

BRITO, Edeleon Marcelo. Mineração de Textos: Detecção automática de sentimentos em comentários nas mídias sociais. 2017. Dissertação de Mestrado Profissional em Sistemas da Informação e Gestão do Conhecimento) — Universidade Fundação Mineira de Educação Cultural. Belo Horizonte, 2017

Avanços nas técnicas de análise automática de documentos, possibilitaram o reconhecimento de aspectos subjetivos em textos extraídos de mídias sociais. O objetivo deste trabalho é implementar um modelo de sistema para classificar automaticamente sentimentos, em bases textuais, escrita em Português do Brasil, utilizando os conceitos da aprendizagem de máquina. O trabalho apresenta, ainda, uma revisão bibliográfica e documental sobre os conceitos gerais da mineração de textos, técnicas de pré-processamento e classificadores de opiniões. Foi também realizado uma revisão sistemática de literatura com o objetivo de estabelecer as principais técnicas, algoritmos e métricas utilizadas na análise de sentimentos. A fim de guiar esta pesquisa, adotou-se o método design science research (DRS). O paradigma DSR é apropriado para orientar a condução de pesquisas científicas em gestão, tecnologia e sistemas de informação. Utilizando a DSR gera-se conhecimento no processo de concepção de artefatos, no caso desta pesquisa, o modelo concebido de análise de sentimentos. Complementa-se esta escolha metodológica com dois experimentos, bem como a utilização do classificador supervisionado Naïve Bayes, implementado a partir do Natural Language ToolKit. Os resultados demonstraram que, no geral, o desempenho do método proposto neste trabalho, para analisar mais de duas polaridades, utilizando a técnica de classificação é inferior às outras ferramentas que testam apenas duas polaridades. Ainda que os resultados não tenha sido expressivos, com relação à acurácia da ferramenta, os achados mostraram que há um grande caminho a percorrer no que diz respeito ao tratamento de dados, especificamente para língua portuguesa. Esse trabalho contribui, então, em grande parte, para viabilização de trabalhos futuros em mineração de textos para a língua portuguesa utilizando módulos de reconhecimento de entidades linguísticas. Espera-se que a pesquisa possa subsidiar propostas de sistemas de tratamento automático de textos publicados em mídias sociais independente do domínio de negócio.

Palavras-chaves: Análise de Sentimentos, Classificação, Polaridade, Opinião.

Abstract

Advances in the techniques of automatic document analysis enabled the subjective aspects recognition in texts extracted from social media. The objective of this work is to implement a system model to automatically classify feelings, in textual bases, written in Brazilian Portuguese language, using the machine learning concepts. The research also presents a bibliographical and documentary review on the general text mining concepts, preprocessing techniques and opinions classifiers. A systematic literature review was also carried out with the objective of establishing the main techniques, algorithms and metrics used in the feelings analysis. In order to guide this research, the design science research (DRS) method was adopted. The DSR paradigm is appropriate to guide the scientific research conduct in management, technology and information systems. Using DSR, knowledge in the artifact design process is generated, in the case of this research, the designed feeling analysis model. This methodological choice is complemented by two experiments, as well as the use of the Naïve Bayes supervised classifier, implemented from the Natural Language ToolKit. The results showed that generally the performance of the method proposed in this work, to analyze more than two polarities, using the classification technique is inferior to the other tools that test only two polarities. Although the results were not expressive, with respect to the tool accuracy, the findings showed that there is a long way ahead to go in regards to data processing, specifically for Portuguese language. This work contributes, in large part, to the feasibility of future work on text mining for the Portuguese language using modules for the linguistic entities recognition. It is hoped that the research may support proposals for systems of automatic treatment of texts published in social media independent of the business domain.

Key-words: Analysis of Sentiments, Classification, Polarity, Opinion.

Lista de ilustrações

Figura 1 – Componentes de um sistema de recuperação de informação	20
Figura 2 – Diagrama que ilustra a metodologia de mineração de textos	26
Figura 3 – Estrutura do classificador <i>Naive Bayes</i> com 5 atributos e uma classe	32
Figura 4 – Características dos trabalhos relacionados	35
Figura 5 – Algoritmos mais utilizados	44
Figura 6 – Métricas de avaliação	46
Figura 7 – Arquitetura em alto nível de um sistema de extração	54
Figura 8 – Sequência de passos para o algoritmo de stemming RSLP	56
Figura 9 – Diagrama de funções do sistema implementado	59
Figura 10 – Diagrama de classes da etapa de pré-processamento.	60
Figura 11 – Diagrama de classes parcial do sistema com a inclusão das classes que compoem o módulo de mineração.	61
Figura 12 – Matriz de confusão	62
Figura 13 – Tela principal da ferramenta	79
Figura 14 – Tela de inclusão da base de dados e Stopwords	80
Figura 15 – Tela para realizar classificação de novas frases	80
Figura 16 – Tela para verificar acurácia	81
Figura 17 – Tela para conferir dados classificados	81

Lista de tabelas

Tabela 1 – Identificação e remoção de <i>stopwords</i> (os <i>tokens</i> descartados estão sublinhados)	29
Tabela 2 – Demonstração do algoritmo de <i>stemming</i>	31
Tabela 3 – Matriz de confusão	33
Tabela 4 – Taxonomia de polaridade e classificação de sentimentos	35
Tabela 5 – Quantidade de itens obtidos pelas buscas e após aplicação dos critérios de inclusão/exclusão	40
Tabela 6 – Sumário de artigos	42
Tabela 7 – Sumário de artigos	43
Tabela 8 – Diretrizes do Design Science Research	49
Tabela 9 – Análise comparativa de mineradores	50
Tabela 10 – Acurácia do classificador para cada uma das categorias testadas	64
Tabela 11 – Acurácia do classificador para cada uma das categorias testadas	65
Tabela 12 – Regras de redução de plural	82
Tabela 13 – Regras de redução de feminino	83
Tabela 14 – Regras de redução advérbio	83
Tabela 15 – Augmentative/diminutive reduction rules	83
Tabela 16 – Regras de redução substantivos	84
Tabela 17 – Regras para redução de verbos	85
Tabela 18 – Regras de remoção de vogal	86

Sumário

1	INTRODUÇÃO	12
1.1	Motivação e Justificativa	13
1.2	Objetivos	14
1.3	Adequação a linha de pesquisa	14
1.4	Estrutura do Texto	15
2	FUNDAMENTAÇÃO TEÓRICA	16
2.1	Fundamentos da mineração de textos	16
2.2	Análise de Sentimentos	17
2.2.1	Terminologia da análise de sentimentos	17
2.3	Áreas de Conhecimento em Mineração de Textos	19
2.3.1	Recuperação de Informação	19
2.3.2	Aprendizagem de Máquina	23
2.3.3	Processamento de Linguagem Natural	23
2.4	Mineração de Textos	25
2.4.1	Classificação de Textos	25
2.5	Etapas da Mineração de Textos	26
2.5.1	Extração	26
2.5.2	Pré-Processamento	27
2.5.3	Mineração	27
2.5.4	Interpretação	27
2.6	Técnicas de Pré-Processamento em Mineração de Textos	27
2.6.1	Tokenização	28
2.6.2	Remoção de <i>stopwords</i>	28
2.6.3	Redução do léxico	29
2.6.4	Normalização de palavras	30
2.6.5	Classificação automática de documentos	31
2.6.6	<i>Naive Bayes</i>	32
2.6.7	Performance de classificadores	33
2.7	Trabalhos Relacionados	34
3	REVISÃO SISTEMÁTICA DE LITERATURA SOBRE TÉCNICAS E MÉTODOS APLICADOS NA ANÁLISE DE SENTIMENTOS	37
3.1	Introdução	37
3.1.1	Análise de Sentimentos	37
3.2	Planejamento	38

3.2.1	Questão de Pesquisa	38
3.2.2	Amplitude da Pergunta	38
3.2.3	Estratégia de Busca	39
3.2.4	CrITÉrios de Seleção	39
3.2.5	EstratÉgias para Extração de Dados	40
3.3	Resultados	40
3.3.1	Realização	40
3.3.2	Questão: Quais técnicas, Domínios e métricas de avaliação estão sendo aplicados na análise de sentimentos?	41
3.3.2.1	Técnicas para classificação de sentimentos	41
3.3.2.2	Domínios	44
3.3.2.3	Métricas	45
3.4	Síntese do capítulo	45
3.5	Considerações finais e trabalhos futuros	46
4	METODOLOGIA	48
4.1	Caracterização da Presente Pesquisa	48
4.2	Diretrizes de Hevner aplicadas a presente pesquisa	48
4.3	Relevância do Problema	49
4.3.1	Comparativos entre as ferramentas comerciais de PLN	50
4.4	Artefatos	52
4.5	Processo de Busca da Solução	53
4.5.1	Ciclo 1 do Design: Implementação	53
4.5.1.1	Aprendizagem automática	53
4.5.1.2	Léxico computacional	53
4.5.1.3	Linguagem <i>Python</i>	54
4.5.1.4	<i>Natural Language Toolkit</i>	55
4.5.1.5	<i>Stemmer</i>	56
4.5.1.6	Algoritmo de <i>Stemmer Portuguese</i>	57
4.5.2	Ciclo 2 do Design: Arquitetura Geral do Sistema	58
4.5.3	Coleta de dados	58
4.5.4	Pré-Processamento	58
4.5.5	Mineração	60
4.5.6	Visualização e análise dos resultados	61
4.5.7	Interfaces e funcionalidades da ferramenta	61
4.6	Rigor da Pesquisa	62
4.7	Avaliação	63
4.7.1	Experimentos	63
4.7.1.1	Experimento 1	63
4.7.1.2	Erros em classificações	64

4.7.1.3 Experimento 2	64
4.7.2 Análise dos experimentos	65
4.8 Síntese das contribuições gerais desta pesquisa	66
4.9 Conclusão	66
4.10 Limitações do trabalho	67
4.11 Trabalhos futuros	69
 REFERÊNCIAS	 71
 Apêndice A	 79
 Apêndice B	 82

1 Introdução

Com o rápido crescimento das mídias sociais tornaram-se visíveis “sentimentos” a respeito dos mais variados assuntos, transformando-se rapidamente em uma verdadeira plataforma de informação e comunicação instantânea, registrando publicamente pensamentos, opiniões, emoções (LIU, 2012).

Instituições, pessoas e empresas estão interessadas em saber qual a opinião de forma automática sobre um grupo de pessoas sobre um determinado tema. Por exemplo, uma universidade pode interessar-se em medir a aceitação de um novo curso, monitorando as opiniões de um grupo em relação a esse tema em um site de mídia social (GONÇALVES et al., 2013).

O uso de mídias sociais tem crescido rapidamente, em 2011, quatro de cinco americanos visitaram sites de mídia sociais e *blogs* (BARNES; LESCAULT, 2011). Números impressionantes que mostraram o aumento da popularidade e a importância dos meios de comunicação social. Além disso, a mídia social tornou-se uma mercadoria neutra em termos de idade, usado tanto por homens e mulheres de todas faixas etárias (STROUD, 2008). Esse crescimento acelerado estimulou estudos e desenvolvimento de sistemas para avaliação de opiniões automaticamente e, conseqüentemente, extração da informações úteis em textos (HAN; KAMBER; PEI, 2011).

Todavia, a tarefa de acompanhar e identificar aspectos importantes para tomada de decisões, diante do grande volume de opiniões expressadas através de postagens por parte dos usuários de mídia sociais, são complexas. Em especial, pela dificuldade de tratá-las por não possuírem um formato de dados estruturados, que, na maioria das vezes, encontra-se em um formato semi-estruturado. Uma vez que esses dados cresceram exponencialmente, e estão disponíveis em diversas plataformas, faz-se necessário utilizar técnicas de recuperação e tratamento, a fim de analisá-los de forma consistente (HAN; KAMBER; PEI, 2011).

Nesse contexto surgiu a área Análise de Sentimentos (AS), também conhecida como Mineração de Opinião, que estuda as opiniões e emoções expressas em textos através das técnicas da inteligência artificial, e na identificação automática da polaridade de opiniões (LIU, 2012). Uma característica importante dessa vertente é classificação de textos de acordo com critérios determinados por um sistema.

Han, Kamber e Pei (2011) afirmam que somente uma pequena parte dos documentos analisados, serão relevantes para um determinado fim. Contudo, sem conhecer o que está contido em cada texto, é praticamente impossível extrair deles quaisquer informação útil. Para isso, foram criados métodos e técnicas (Seção 2.6) para analisarem documentos e classificá-los de acordo com rótulos pré-estabelecidos.

1.1 Motivação e Justificativa

A Análise de Sentimentos, tem sido escolhida pelo mundo acadêmico como uma linha útil de pesquisa, e recebe um interesse cada vez maior da comunidade de processamento de linguagem natural. Esse interesse crescente é particularmente motivado pela necessidade generalizada de aplicações baseadas em opiniões, tais como análises de produtos e comentários de filmes, rastreamento de entidades e análises de sumários.

O uso das técnicas relacionadas a essa área permite recuperar informações, extrair dados, resumir documentos, descobrir padrões, associações, regras e realizar análises qualitativas e quantitativas em documentos de texto. A partir desta perspectiva, surgiram diversas frentes de pesquisas com o aperfeiçoamento de técnicas voltadas para a seleção de características [Pak e Paroubek \(2010\)](#) e a identificação de diferentes tópicos [\(YU; LIU, 2004\)](#).

Quanto ao idioma, no cenário internacional das pesquisas em computação, a análise de sentimentos, através do processamento de linguagem natural faz parte da agenda de pesquisa das mais importantes universidades no mundo. Contudo, além do muito que ainda existe por ser feito, as pesquisas em mineração de textos remetem a uma forte regionalização, apresentando maiores resultados para algumas línguas do que para outras. [\(HAN; KAMBER; PEI, 2011\)](#)

De contornos deliameados, ao longo do tempo, por aspectos sensivelmente econômicos, os resultados destas pesquisas se voltam ainda timidamente à língua portuguesa. Sendo concentrado na língua inglesa a grande parte dos produtos gerados, e estes não são, na maioria das vezes transportáveis para outras línguas. [\(JACKSON; MOULINIER, 2007\)](#)

Além disso, instituições de ensino, empresas públicas e privadas lidam diariamente com um grande número de dados na Internet, tais como notícias, legislações, relatórios entre outros tipos de documentos produzidos internamente. Esse panorama justifica as pesquisas em processamento e análise de sentimentos na língua portuguesa. Duas pesquisas podem comprovar essa linha de raciocínio. A primeira, descrita por [Tan et al. \(1999\)](#) mostrou que 80% das informações armazenadas nas empresas são também dados não-estruturados. A segunda mostrou que 80% do conteúdo contido na Internet esta em formato textual [\(CHEN, 2001\)](#). Por esses e outros motivos, a análise automática de textos se torna importante para quem lida com informação e conhecimento.

Portanto, é neste cenário que se assenta este trabalho, como problema de pesquisa, surge a seguinte indagação:

Como os atuais métodos de classificação de sentimentos, que utilizam de características sintáticas, poderão classificar automaticamente um texto escrito em português do Brasil, extraído de mídias sociais.

Este problema pode ser sumarizado na forma da seguinte questão de pesquisa:

"Um sistema de mineração de textos, baseado nas técnicas de classificação, poderá obter bons resultados classificando um texto escrito em Português do Brasil, utilizando as polaridades: Angustiado, Animado, Cansado, Entendiado, Encantado, Feliz, Irritado, Sonolento, Sereno, Satisfeito, Triste e Tenso "

Adicionando ao sistema à característica de classificar mais de duas polaridades, espera-se obter resultados semelhantes ou superiores em comparação aos sistemas sem essa característica.

Para responder as indagações acima citadas é que este trabalho será desenvolvido. Portanto, apresentam-se os objetivos a seguir.

1.2 Objetivos

O objetivo geral dessa dissertação é implementar um modelo de sistema para classificar automaticamente sentimentos, em bases textuais, escrita em Português do Brasil, utilizando os conceitos da aprendizagem de máquina.

Para alcançar este objetivo, delineiam-se os seguintes objetivos específicos:

- Identificar métodos e técnicas frequentemente utilizados na análise de sentimentos;
- Prover um modelo de um sistema de informação para analisar sentimentos em textos;
- Testar o modelo proposto sobre dados coletados através de um site de mídia social;

1.3 Adequação a linha de pesquisa

O programa de Mestrado em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC possui caráter profissional, ou seja, têm foco na realização de pesquisas aplicadas e práticas em áreas gerenciais e tecnológicas e, estrutura-se em uma única área de concentração: Gestão de Sistemas de Informação e do Conhecimento que, encontra-se em duas linhas de pesquisa: Gestão da Informação e do Conhecimento e Tecnologia e Sistemas de Informação.

Este trabalho posiciona-se na linha de pesquisa **Tecnologia e Sistemas de Informação**, por visar o desenvolvimento de sistemas de informação. Nessa linha de pesquisa, é possível enxergar um software que auxilie cientistas da computação em estudos sobre análise e classificação de sentimentos.

Como objetivo, propomos uma ferramenta de detecção automática de sentimentos para assistir outros pesquisadores em estudos sobre análise de sentimentos.

Uma das principais contribuições deste trabalho é a proposta de uma metodologia prática para extrair informações de dados não-estruturados, através de uma ferramenta inteligente. Essa ferramenta terá como objetivo principal a avaliação de frases através de suas características de polaridades.

1.4 Estrutura do Texto

Para melhor organização dessa dissertação, seu conteúdo foi dividido em três capítulos. No Capítulo 1, são apresentadas a motivação e os objetivos geral e específicos do trabalho.

O Referencial Teórico e a revisão sistemática de literatura encontra-se no Capítulo 2 e 3.

O Capítulo 4 apresenta a metodologia, através do método *DSR*, são abordados os conceitos, técnicas e o desenvolvimento da ferramenta. Por fim, ainda no Capítulo 4 são apresentado os resultados alcançados e discutidos.

As referências bibliográficas utilizadas na composição da presente pesquisa são apresentadas ao término do documento.

2 Fundamentação Teórica

Neste Capítulo, são apresentados os principais conceitos que fundamentam a presente pesquisa. Na Seção 2.1 são abordados os conceitos introdutórios que norteiam uma aplicação para análise de sentimentos.

O cenário inicial do trabalho refere-se à implementação de um sistema de informação capaz de classificar sentimentos em base textuais, escrito em Português do Brasil. Desta forma, a Seção 2.5 conceitua as etapas importantes para implementar um sistema desse gênero.

Na sequência, o tema central é retomado. Nas Seções 2.6, 2.6.2, 2.6.3, 2.6.6, são abordadas, respectivamente, as principais técnicas aplicáveis nesse contexto, para que seja alcançado os objetivos da análise de sentimentos, arrematando o tema.

2.1 Fundamentos da mineração de textos

Com o avanço das tecnologias de informação, viu-se um número crescente de aplicações que armazenam dados não estruturados se proliferarem. Estes dados, na maioria das vezes, incluem informações valiosas, como exemplo: tendências, anomalias e padrões de comportamento que podem ser usados para auxiliar nas tomadas de decisões (BERRY; KOGAN, 2010). Por outro lado, com o aumento da utilização de equipamentos digitais, sobretudo aqueles que utilizam da internet como meio de comunicação, fez com que a procura por sistemas capazes de descobrir conhecimento aumentasse exponencialmente (BERRY; KOGAN, 2010). Diante disso, diversas técnicas foram desenvolvidas com propósito de recuperar informações importantes contidas em bases de dados, dando origem à área chamada de *Text Mining*, que deriva das técnicas de *Data Mining*, uma vez que as duas procuram extrair informações úteis em dados não estruturados ou semiestruturados e, estes serem difíceis de tratar (FELDMAN; SANGER, 2007).

Para além das técnicas empregadas no *Data Mining* que são reutilizadas no *Text Mining*, outras áreas de estudos são igualmente importantes para a extração de conhecimento em bases textuais, como exemplo: Aprendizagem de Máquina, Inteligência Computacional (IC), Recuperação da Informação (RI), Ciência Cognitiva e, não menos importante, o Processamento de Linguagem Natural (PLN), que explora como os computadores podem ser utilizados para compreender a linguagem natural (CHOWDHURY, 2003). Existem ainda duas abordagens para o processo de mineração em bases textuais, são elas: a **Análise Estatística**, que trabalha diretamente com a frequência de aparição de cada termo em uma frase não se preocupando com o contexto inserido e a **Análise Semântica**, que se preocupa com a funcionalidade dos termos, através do significado morfológico, sintático, semântico, pragmático, conforme o ser humano faz. Ambas as abordagens podem ser utilizadas sozinhas ou em conjunto (CHOWDHURY, 2003).

Ainda que seja importante contextualizar essas áreas, por serem parte integrante do processo, o foco deste trabalho fundamenta-se na construção de uma ferramenta capaz de classificar polaridade em textos, triste, feliz, tenso, entendiado, alegre entre outros. Desse modo, a contextualização sobre as demais áreas, descritas em capítulos, será mais superficial do que a feita para os métodos e técnicas utilizadas diretamente na Análise de Sentimentos (AS).

2.2 Análise de Sentimentos

A análise de Sentimentos, também conhecida como Mineração de Opinião, corresponde ao problema de identificar emoções, opiniões em textos e, embora tenham tido recentemente uma explosão de estudos nessa área, devido ao seu potencial de aplicabilidade em diversas áreas da ciência, o interesse já existe há algum tempo, de acordo com [Wilson, Wiebe e Hwa \(2006\)](#), [Wiebe et al. \(2004\)](#), [Liu \(2012\)](#) entre outros. Segundo [Pang e Lee \(2008\)](#) foi no ano de 2001 que marcou uma nova era sobre as oportunidades e problemas de investigação que essa área pode nos trazer.

É importante ressaltar que a AS trata de problemas de classificação e que, como tal, é utilizada para classificar textos de acordo como a sua polaridade, mesmo que uma frase não denote explicitamente um sentimento. A frase "Jovem com suspeita de Dengue morre em hospital" apenas descreve um fato, no entanto, poderá ser classificada como positiva ou negativa para a área da saúde ([LI; LEE; HUANG, 2013](#)).

Encontra-se na literatura vários estudos em que a análise de sentimentos foram aplicadas, desde mineração de opinião sobre produtos e serviços [Giatsoglou et al. \(2017\)](#), [Dave, Lawrence e Pennock \(2003\)](#), [Hu e Liu \(2004\)](#), [Pang, Lee e Vaithyanathan \(2002\)](#), [Turney \(2002\)](#); em notícias [Godbole, Srinivasaiah e Skiena \(2007\)](#) e em áreas relacionadas à saúde [Goeuriot et al. \(2012\)](#) e política ([CHEN, 2001](#)). Enquanto alguns trabalhos focam na identificação de opiniões [Wiebe e Riloff \(2005\)](#), [Wiebe et al. \(2004\)](#), [Wilson et al. \(2005\)](#), ou seja, se determinado documento contém opiniões ou simplesmente fatos, outros trabalhos consideram apenas a classificação da polaridade das opiniões encontradas em fóruns, blogs de discussão, sites de comércio eletrônico ([ABBASI; CHEN; SALEM, 2008](#)). Outros transpõem à atividade de classificação de sentimentos, como o trabalho de [Pang e Lee \(2008\)](#) que verifica a detecção de subjetividade, ou seja, se apenas uma parte de um texto possui conteúdo opinativo; na identificação de pontos de vista com [Wiebe et al. \(2004\)](#), na sumarização de opiniões [Hu e Liu \(2004\)](#) e ainda em sistemas de questionários ([STOYANOV; CARDIE; WIEBE, 2005](#)).

2.2.1 Terminologia da análise de sentimentos

A idéia de opinião ou sentimento trabalhada por pesquisadores na análise de sentimentos é ampla e possui algumas terminologias para delimitar a área. De modo a ficar mais claro o conteúdo em torno de dessa área, serão mostrados aqui alguns conceitos.

Segundo [Wiebe et al. \(2004\)](#), a análise de sentimentos trata da detecção automática dos estados privados, que são aqueles que não podem ser observados por outros. Para eles, estes conceitos estão intimamente relacionados a noção de estado interno.

Em outra linha, [Roman \(2007\)](#) diz que antes de falar sobre sentimentos em texto, deve-se trabalhar com análise de emoções. Para este autor, "pode-se distinguir emoções de sentimentos, sendo as segundas as justaposições das alterações no estado corpóreo justaposto à imagem mental do que ocasionou tal mudança, o sentimento consiste em sentir uma emoção".

Por outro lado, [Liu \(2012\)](#), defende a idéia em que uma opinião advém de uma atitude, expressada por um determinado termo polarizado e, associado à um aspecto ou atributo de uma entidade por um indivíduo. "Uma opinião é entao, por natureza, relacional, pessoal e explícita". Ainda este autor, distingue-se as opiniões em dois tipos: diretas e as comparativas. A primeira associa-se diretamente com uma emoção ou atitude de uma entidade; enquanto as comparativas, expressam uma relação de similaridades em dois ou mais objetos.

Ainda [Liu \(2012\)](#) reporta a análise de sentimentos como um conjunto de termos, sendo eles: Objetos, Componentes, Opinião, Polaridade. O objeto é o alvo de análise, pode referir-se a um produto, serviço, pessoa ou uma entidade. O componente refere-se as características do objeto, ou seja, uma opinião pode ser dada sobre um determinado produto, mas ao mesmo tempo sobre uma característica do mesmo. A opinião é a expressão, atitude ou emoção emitida por alguma entidade e, por último, a polaridade, que determina se a opinião é positiva, negativa ou neutra.

Outra importante definição que este autor faz, é que a Análise de Sentimentos pode ainda ser tratada em dois níveis distintos, e com objetivos diferentes:

- Classificação em Nível de Documento: A opinião está em torno de todo texto ou sentença em questão. A opinião é dada como positiva, negativa ou neutra, observando o texto ou sentença por completo.
- Classificação Baseada em Aspectos: Ao invés de classificar o texto como um todo, cada aspecto é classificado.

Na classificação em nível de documento, o resultado da opinião é conseguido através da revisão do texto completo, podendo ser positiva, negativa ou neutra. Contudo, informações podem ser perdidas, visto que dentro de um texto há tanto informações positivas e negativas. Já na classificação baseada em aspectos, nenhuma opinião é descartada, sendo considerada todas mostradas no texto.

O texto abaixo servirá de base para exemplificar este conceito:

"(1) Ontem comemoramos o aniversário da minha tia. (2) A comida no restaurante é maravilhosa,(2.1) e não é caro. (3)A carne oferecida é muito boa, (3.1) mas a batata frita não é das melhores."

Ao analisarmos o exemplo acima, o fator predominante de análise é a subjetividade do texto, ou seja, trata-se de um texto opinativo ou de um texto subjetivo. Sendo este um texto opinativo, extraímos as opiniões para serem analisadas. Analisando a oração (1), percebemos que se trata de um texto subjetivo, por demonstrar um fato ocorrido e não uma opinião acerca de um objeto, com isso a oração seria descartada da AS. Já as orações (2) e (3), demonstram claramente uma opinião. Portanto, para a análise de sentimentos, é importante verificar anteriormente o tipo de classificação a ser usada, pois o resultado desejável pode ser diferente para o usuário final.

No exemplo, a sentença poderia ser classificado como positivo, se analisado pela classificação baseada em documento, considerando que a maioria das opiniões são positivas, mas parte da informação seria perdida. Além disso, na oração (3) há duas opiniões contrárias para o mesmo objeto, porém para aspectos diferentes. Na classificação baseada em texto, essa informação será desconsiderada, enquanto que na classificação baseada em aspecto, não.

Contudo, nem sempre é possível classificar as sentenças, sendo primeiramente e importante separar as sentenças em classificáveis e não classificáveis (LIU, 2012). Abaixo lista-se as definições usadas para esses dois tipos de sentenças:

- Sentença Objetiva: É aquela que não possui a opinião do autor, é mostrado apenas alguns fatos sobre o objeto em questão.
- Sentença Subjetiva: É aquela que apresenta uma opinião ou crença do autor a respeito de determinado objeto.

Seguindo o nosso exemplo, na oração (1) temos uma sentença objetiva, enquanto que nas orações (2) e (3) apresentam sentenças subjetivas.

Como o objetivo de atingir melhores resultados, a análise de sentimentos em textos é dividida em tarefas que normalmente são sequenciais e complementares, ver (Seção 2.6) dado a complexidade de cada uma dessas tarefas

2.3 Áreas de Conhecimento em Mineração de Textos

Na seção 2.1 foram apresentados os fundamentos acerca da mineração de textos e sobre a análise de sentimentos. A presente seção visa apresentar forma resumida, porém suficiente para o entendimento, conceitos importantes utilizados neste trabalho.

2.3.1 Recuperação de Informação

Recuperação de Informação é um processo de seleção de documentos que atende os requisitos de usuários Nasukawa e Nagano (2001) e respondem à necessidade de informação com o auxílio de índices Ananiadou e McNaught (2006). Os sistemas tradicionais de RI usualmente aplicam

termos de índice para indexar e recuperar documentos [Baeza-Yates \(1999\)](#). Os sistemas detectam e extraem documentos de interesse, combinando a consulta dada [Feldman e Sanger \(2007\)](#), a partir da enorme quantidade de documentos [Nasukawa e Nagano \(2001\)](#) e os apresentam ao usuário ([CHOWDHURY, 2003](#)). Mas exige que o usuário leia os documentos para localizar as informações relevantes [Feldman e Sanger \(2007\)](#), uma vez que os usuários não têm uma clara intenção do que precisam [Nasukawa e Nagano \(2001\)](#).

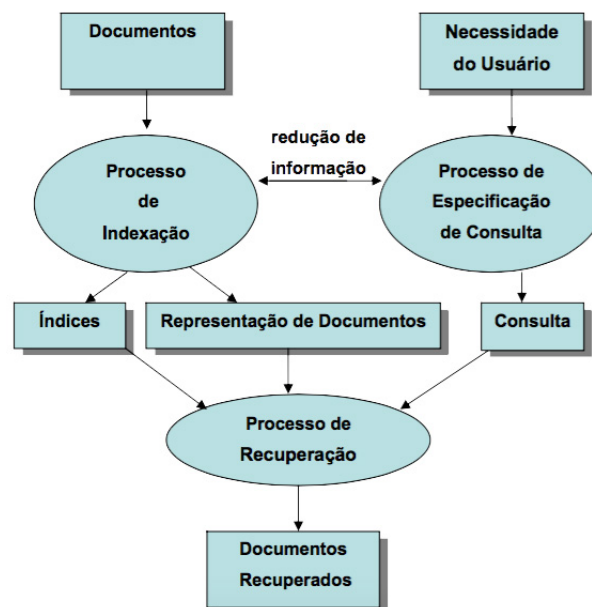
[MOOERS \(1951\)](#) cunhou o termo recuperação da informação, destacando que ele "engloba os aspectos intelectuais da descrição de informações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregados para o desempenho da operação".

O processo de recuperação da informação baseia-se em técnicas tradicionais da ciência da computação e, em recursos óbvios dos dados para criar estruturas de índices, organizar e recuperar de forma eficiente as informações. Essa estrutura de índice permite identificar, no conjunto de documentos (corpus) de um sistema, quais atendem à necessidade de informação do usuário.

As técnicas de recuperação da informação estão intimamente relacionada com a mineração de textos, principalmente no processo de **Indexação**, em que são montados filtros para eliminar palavras de pouca significação (stop words), além de normalizar os termos reduzindo-os a seus radicais, processo conhecido como **stemming** (2.6.4).

Um sistema tradicional de Recuperação de Informação pode ser estruturado conforme ilustrado na Figura 1.

Figura 1 – Componentes de um sistema de recuperação de informação



Fonte adaptada: [Baeza-Yates \(1999\)](#)

O **Processo de Indexação** cria estruturas de dados ligadas à parte textual dos documentos, por exemplo, as listas invertidas. As listas invertidas são ordenadas por tipo <chave-valor>, em que as chaves são termos do vocabulário da coleção e os valores são listas de referências para documentos.

O **Processo de Especificação da Busca** na maioria dos casos é uma tarefa difícil. "Há frequentemente uma distância semântica entre a real necessidade do usuário e o que ele expressa na consulta formulada". Essa distância é gerada pelo limitado conhecimento do usuário sobre o universo de pesquisa e pelo formalismo da linguagem de consulta (BAEZA-YATES, 1999).

O **Processo de Recuperação** gera uma lista de documentos respondendo à consulta formulada pelo usuário. Os índices especificados para aquela coleção de documentos são utilizados para acelerar o procedimento.

Com a abordagem de correspondência de palavras, o problema é às vezes simplificado demais, uma vez que as consultas dos usuários podem ser perdidas quando sua intenção é substituída por um conjunto de palavras (BAEZA-YATES, 1999). Além disso, as consultas de correspondência com os termos de índice, representando documentos, podem fazer com que os documentos recuperados sejam irrelevantes para as solicitações. Por conseguinte, uma das principais preocupações no IR é distinguir entre documentos relevantes e irrelevantes. Algoritmos de classificação foram aplicados para auxiliar essa decisão. Os algoritmos estabelecem uma ordenação simples dos documentos recuperados, de acordo com a relevância do documento. Modelos de infravermelho foram projetados, com base nessas características, para servir os propósitos de RI (FELDMAN; SANGER, 2007).

Embora esse seja um campo vasto, neste trabalho o interesse em RI se restringe à representação e identificação de documentos sobre conjuntos de assuntos específicos.

Pode-se considerar que a recuperação da informação é a primeira etapa para o processo de mineração de textos. Um sistema RI executa filtragens de documentos sobre um conjunto de dados, mostrando ao usuário o resultado de uma consulta em particular. Na literatura, as técnicas de RI são separadas em dois grupos: tradicional e moderno. Os modelos tradicionais são comumente utilizados para Recuperação de Informação e os mais conhecidos são:

1. (modelos booleanos)
2. (vetoriais)

Modelo booleano: É um modelo simples baseado na teoria de conjuntos e álgebra booleana. As solicitações são representadas como expressões booleanas com semântica precisa. Suas vantagens são a simplicidade e o formalismo limpo por trás do modelo, enquanto a principal desvantagem é a correspondência exata que pode levar a muitos ou muito poucos documentos recuperados. No modelo booleano, a semântica da consulta é bem definida, em que, cada documento corresponde à

expressão booleana, ou não. Em função dessa simplicidade, esse modelo é amplamente utilizado em ferramentas de buscas comerciais, muito devido a sua descomplicada semântica e do cálculo direto dos resultados utilizando operações de conjunto (ANANIADOU; MCNAUGHT, 2006).

Nesse modelo, para que o usuário possa expressar suas necessidades de informação é necessário algumas habilidades técnicas. É importante que o usuário esteja familiarizado com as premissas booleanas, no caso, as *queries*, também conhecidas como consultas complexas. Grande parte dos mecanismo de busca utilizam das *queries* (O'RIORDAN; SORENSEN, 1997).

Ainda que o modelo booleano seja bastante difundido, problemas são encontrados:

- Se o usuário não conhecer bem o domínio de negócio, a seleção dos termos, através da formulação de consultas adequadas torna-se complexa.
- O tamanho do resultado da consulta não pode ser controlado. Os dados resultante pode conter nenhum ou vários itens. Além disso, sem conhecimento técnico de *queries* o grau de comparação pode ser prejudicado, sendo difícil saber o que foi deixado de fora na definição da consulta.

Modelo de Espaço Vetorial(VSM): É um modelo que tenta resolver problemas de representação de documentos utilizando representação geométrica. Sendo muito utilizado, também, em tarefas que é necessário encontrar documentos que atendam à uma especificidade, e a solução desse problema desenrola-se espontaneamente do esquema de representação dos documentos. Documentos são caracterizados através de pontos (ou vetores) em um espaço Euclidiano dimensional em que cada dimensão condiz a uma palavra (termo) do vocabulário (SALTON, 1989). Cada palavra tem um peso associado para retratar sua significância, que pode ser sua frequência em um documento. A semelhança entre dois documentos é definida através da distância entre os pontos ou com o ângulo entre os vetores, neste caso, desconsidera-se o comprimento do documento. Isso acontece para que seja levado em conta os documentos de tamanhos diferente. Assim, é possível normalizar os documentos de forma que fique com um comprimento unitário (BAEZA-YATES, 1999).

Por se tratar de um modelo simplista, o VSM utiliza-se de uma forma bastante intuitiva para representar documentos textuais, possuindo algoritmos padronizados para realizar a redução da dimensão, a seleção de características e a visualização de espaços de vetores de forma rápida e eficiente. (BAEZA-YATES, 1999). Uma característica negativa nesse modelo é a alta dimensionalidade: a quantidade de palavras diferentes em uma coleção de documentos atinge facilmente milhares de registros. Outro problema conhecido é composto pelos erros de grafia e estilo de escrita.

A representação binária nem sempre atende as necessidades de determinados usuários, pois em muitos casos deve ser utilizada uma medida levando em consideração a frequência em que um termo aparece no documento. Para estes casos, usa-se os modelos de atribuição de pesos.

Modelo Atribuição de Pesos

Nesse modelo, os documentos são representados por vetores, em que as dimensões são os termos presentes no corpus inicial a ser minerado. Cada coordenada no vetor é um termo que possui um valor numérico com sua relevância para o documento. O processo de associar valores numéricos a coordenadas de vetor é descrito na literatura como atribuição de pesos ou, simplesmente, *weighting*. Sendo este, um processo de dar ênfase aos termos mais importantes. Algumas medidas mais populares para atribuição de pesos *weighting* são: Binária e TF. No binário usa-se os valores 0 e 1 para divulgar se um termo existe em um documento ou não, respectivamente. A TF (frequência do termo) faz uma contagem das ocorrências de um termo no documento usando um contador como medida numérica, normalizando essas medidas para valores no intervalo [0,1]. Sem a normalização, um termo pode ter uma medida maior em um determinado vetor-documento. Por exemplo, considere dois documentos que pertencem à mesma categoria com tamanhos de 10 *Kbytes* e 100 *Kbytes*. Nesse caso, haveria uma grande diferença na frequência dos termos em ambos os documentos. Então, a normalização é necessária para ajudar a resolver problemas em que o comprimento do documento é muito grande (SALTON, 1989).

2.3.2 Aprendizagem de Máquina

Aprendizado de Máquina estuda a criação de modelos algoritmos probabilísticos capazes de “aprender” através da experiência. O aprendizado se dá através de métodos dedutivos para extração de padrões em grandes massas de dados Chakrabarti (2002). *text Machine Learning* (ML) tem sido muito utilizado no processo de classificação automática de textos.

Segundo Mitchell (1997) o aprendizado de máquina estuda como os algoritmos computacionais são capazes de automaticamente melhorarem a execução de tarefas através da experiência. Os algoritmos desenvolvidos sobre a aprendizagem de máquina se baseiam na estatística e probabilidade para aprender padrões complexos a partir de algum corpus.

Na literatura encontramos diversos trabalhos interdisciplinares que aplicaram os algoritmos de aprendizado de máquina. Com exemplo, o trabalho de correção ortográfica Schmid (1994) e o de diagnóstico de doenças Bair e Tibshirani (2003). A aplicabilidade desse algoritmo rendeu bons resultados, muito devido a sua utilização em diversas áreas de natureza diferenciadas.

Os casos de sucesso do aprendizado de máquina nos leva a acreditar que esse tipo de algoritmo é fortemente relacionado à área de mineração de textos, estendendo-se a classificação automática de textos.

2.3.3 Processamento de Linguagem Natural

Devido a escalabilidade dos gerenciadores de bancos de dados em armazenar informações, diversos sistemas conseguiram manter disponíveis textos em formato de documentos sem problemas de demanda, acesso e disponibilidade dos dados. Todavia, com o aumento exponencial

de documentos circulando em diversos tipos de sistemas, mesmo os computadores modernos podem não comportar essa massa de dados, tendo que restringir a representação à um conjunto limitado de termos. Além disso, o que os usuários necessitam é representado por uma expressão de busca, que pode ser especificada em linguagem natural. Mas, esse mecanismo de expressão de busca traz dificuldades para a maioria dos usuários, pois eles têm que prever as palavras ou expressões que satisfaçam sua necessidade (LIDDY, 2001).

O Processamento de Linguagem Natural (PLN) surge então, para resolver problemas relacionados à recuperação da informação, ao observar que os documentos e as expressões de busca são apenas objetos linguísticos. Através dessa observação, criou-se várias técnicas dentro da PLN para analisar textos em um ou mais níveis linguísticos, com intuito de emular o processamento humano da língua (LIDDY, 2001).

O PLN é uma área de Ciência da Computação que estuda como os computadores podem analisar e/ou gerar textos em linguagem natural Perna, Delgado e Finatto (2010). Turban et al. (2010) descreve que o processamento da linguagem natural pode ser vista como a forma de comunicação entre o homem e a máquina, sendo essa comunicação em qualquer linguagem que se fale. Os autores ainda dizem que:

Para entender uma consulta em linguagem natural, o computador precisa ter conhecimento para analisar e interpretar a entrada de informação. Isso pode significar conhecimento linguístico de palavras, conhecimento sobre áreas específicas, conhecimentos gerais e até mesmo conhecimento sobre os usuários e seus objetivos. No momento em que o computador entende a informação, ele pode agir da forma desejada (TURBAN et al., 2010).

Para Lopes (2002) O PLN não é uma tarefa trivial devida a natureza ambígua da linguagem natural. Essa diversidade faz com que o PLN difere do processamento das linguagens de programação de computador, as quais são fortemente definidas para evitar a ambiguidade.

Ainda este autor classifica as técnicas de PLN conforme o nível linguístico processado: fonológico, morfológico, lexical, sintático, semântico e pragmático. Estes níveis precisam ser entendidos e diferenciados. Especificamente, o morfológico que trata das palavras isoladamente, o léxico que trabalha com o significado das palavras, o sintático que se refere a estrutura das frases, o fonológico que lida com a pronúncia, o semântico que interpreta os significados das frases (LIDDY, 2001).

Todas essas técnicas podem ser utilizadas em um processo de PLN, contudo, para o presente trabalho, o nível fundamental é o morfológico. O analisador morfológico tem o propósito de selecionar as palavras e expressões que estão isoladas no texto.

É importante ressaltar que existem técnicas dentro da PLN que não são aplicáveis a Mineração de Textos, como exemplo, as correções ortográficas e a tradução automática de textos (JUNIOR, 2007).

2.4 Mineração de Textos

O principal objetivo da Mineração de Textos (MT) consiste na extração de características em uma grande quantidade de dados não estruturados. Segundo [Weiss et al. \(2010\)](#) as técnicas utilizadas na mineração de textos são semelhantes as utilizadas na mineração de dados, ou seja, fazem o uso dos mesmos métodos de aprendizagem, independente se uma técnica utiliza-se de dados textuais (MT) e a outra com dados numéricos (MD).

Pode-se diferenciar as duas técnicas a partir de dois conceitos: enquanto a Mineração de Dados é caracterizada por extrair informações implícitas, anteriormente desconhecidas, contudo potencialmente úteis. Na Mineração de Textos, a informação que se deseja extrair é clara, sendo explicitada nos textos, porém o problema é que a informação não é expressa de uma maneira que seja passível de processamento automático ([WITTEN; FRANK, 2011](#)).

De fato, estamos vivenciando o crescimento acelerado de informações não estruturadas (textos), com isso, a Mineração de Textos ganha espaço não somente no meio acadêmico, mas também no mundo dos negócios.

Resumidamente, a área de estudo da MT compreende cinco tarefas triviais ao processo: Recuperação de Informação, Pré-Processamento de Textos, Sumarização, Classificação Automática de Textos e Análise dos dados.

2.4.1 Classificação de Textos

Um dos motivos para o crescente interesse no estudos sobre a área da mineração de textos, especificamente na técnica de classificação, é devido ao crescimento e a disponibilidade de documentos na internet, sobretudo pelas redes sociais.

A técnica dominante para este problema é baseada na aprendizagem de máquina, ou seja, um processo indutivo cria automaticamente um classificador por “aprendizado”, a partir de um conjunto de dados classificados previamente. A vantagem dessa abordagem é a independência de domínio. Pode-se dizer então que a tarefa de classificar um texto automaticamente é uma derivação da aprendizagem de máquina com o propósito de atribuir rótulos pré-definidos a documentos textuais ([SEBASTIANI, 2002](#)).

[Sebastiani \(2002\)](#) assegura que a classificação de textos consiste em determinar se um documento d_i , (de um conjunto de documentos D) é pertencente ou não a uma categoria c_j (de um conjunto de categorias C), consistentemente com o conhecimento das categorias corretas para um conjunto de documentos de treinamento.

O objetivo principal da classificação é atribuir uma determinada classe à uma conjunto de documentos e, no caso da análise de sentimentos, trata-se de classificar automaticamente um conjunto de dados às classes positivas e negativas.

2.5 Etapas da Mineração de Textos

Neste capítulo apresentaremos a metodologia proposta por [Aranha e Vellasco \(2007\)](#) para Mineração de Textos. Em seu trabalho, Aranha descreve como sendo um modelo completo para adquirir conhecimentos a partir de um corpus textual. O objetivo deste capítulo é detalhar todas as etapas e técnicas desta metodologia, uma vez que este processo é o que melhor se enquadra no presente trabalho. A figura 2 ilustra a metodologia.

Figura 2 – Diagrama que ilustra a metodologia de mineração de textos



Fonte: Adaptado de [Aranha e Vellasco \(2007\)](#)

[Aranha e Vellasco \(2007\)](#) descrevem em seu trabalho a metodologia dividida em cinco etapas distintas, a primeira na coleta dos dados, a segunda no pré-processamento dos mesmos, com intuito de criar o primeiro nível de estruturação, a terceira etapa confere a criação dos índices que possibilitam uma melhora na recuperação dos dados, a quarta na aquisição de conhecimento e, por fim, uma quinta fase para interpretação dos resultados obtidos.

2.5.1 Extração

Na mineração de textos, quando estamos diante de um problema de classificação automática de documentos, faz-se necessário obter um conjunto de dados para treinamento [Aranha e Vellasco \(2007\)](#). Portanto, a etapa de extração e coleta de dados tem como função a criação de uma base de dados textual.

Segundo [Manning et al. \(2008\)](#) a coleta poderá ser realizada utilizando-se de *crawlers*. *Crawler* é um *software* que percorre sítios da *internet* com intuito de coletar automaticamente os dados destes. Após a recuperação destes dados pretendidos para a análise, é possível criar um corpus que servirá de base para aplicar as técnicas de mineração de textos.

Um corpus nada mais é que uma coleção de textos, que representa uma ou um conjunto de linguagens naturais e, a criação deste conjunto de treino revela-se uma tarefa custosa, uma vez que na maioria dos casos exige-se processos manuais à base expert judgment ([INDURKHYA; DAMERAU, 2010](#)).

2.5.2 Pré-Processamento

Pré-processamento é a etapa executada imediatamente após a coleta dos dados. Pré-processar textos é, na maioria das vezes, uma etapa muito onerosa, uma vez que utiliza-se diversos algoritmos que consomem boa parte do tempo do processo de extração de conhecimento e, por não existir uma única técnica que possa ser aplicada em todos os domínios de aplicações.

O principal objetivo de pré-processar um texto, consiste na filtragem e limpeza dos dados, eliminando redundâncias e informações desnecessárias para o conhecimento que se deseja extrair (GONÇALVES et al., 2006). As principais técnicas aplicadas nesta etapa, considerando bases de dados textuais, são apresentadas na seção 2.6.

2.5.3 Mineração

Na etapa de Mineração são aplicadas técnicas direcionadas ao aprendizado de máquina para obtenção de novos conhecimentos Witten e Frank (2011). Nessa etapa escolhemos qual tarefa de acordo com a necessidade do usuário. Por exemplo, se a necessidade for verificar o grau de similaridade e a formação de grupos naturais, então a tarefa a ser escolhida é **clusterização**. Por outro lado, se estes grupos já estão formados, seja por conhecimento prévio do especialista ou pela execução de algoritmos, então a orientação de aonde um novo documento deve ser “rotulado” é conseguida através de algoritmos de **classificação**.

No contexto deste trabalho, as técnicas aplicadas na etapa de mineração, devem ser capazes de identificar as características que diferenciam documentos pertencentes a diferentes classes e, realizar o processo de classificação.

2.5.4 Interpretação

A etapa da interpretação dos dados é onde será validada a eficiência do processo como um todo, analisando os dados obtidos após aplicação dos algoritmos na etapa anterior. Em outras palavras, é nesta etapa que avaliamos se o objetivo de descobrir um novo conhecimento foi adquirido, a partir de uma base de dados (WITTEN; FRANK, 2011).

Por fim, vale ressaltar que este processo é cíclico. Ao final de cada uma das etapas, os resultados devem ser analisados individualmente e caso não apresentem resultados satisfatórios, deve-se realizar alterações no processo para a realização de um novo ciclo.

2.6 Técnicas de Pré-Processamento em Mineração de Textos

Nesta seção apresentaremos as técnicas que serão utilizadas no presente trabalho, na etapa de **Pré-Processamento**, com intuito de alcançar melhores resultados no refinamento de informações.

Quando trabalha-se com base textuais, existe uma grande quantidade de termos e atributos para sua representação, resultando assim, em uma denotação esparsa, em que grande parte dos atributos são nulos. Dessa forma, as técnicas aplicadas no Pré-Processamento são importantes para resolver problemas em que dados textuais estão envolvidos. Portanto, uma boa amostragem dos dados é aquela que, identifica os melhores atributos que representam o conhecimento e, que consiga reduzir drasticamente a quantidade destes sem perder as características principais da base de dados.

2.6.1 Tokenização

A identificação de *tokens*, ou tokenização é uma importante etapa do Pré-Processamento para extrair unidades mínimas de textos. Cada unidade é chamada de token e, normalmente, corresponde à uma palavra do texto, podendo estar relacionado também a símbolos e caracteres de pontuação, com exemplo “ ”, “.”, “! (MANNING et al., 2008).

O termo *Token* será bastante utilizado nessa dissertação, visto que, em alguns momentos, ele poderá possuir o mesmo sentido de “palavra”. De fato, na maioria das vezes um *token* representa uma palavra no corpus. Como exemplo, a frase: “Amanhã iremos para Belo Horizonte!”, esta frase poderá ser dividida em seis tokens. Conforme mostra o exemplo abaixo.

“Amanhã iremos para Belo Horizonte!”

[Amanhã] [iremos] [para] [Belo] [Horizonte] [!]

Na geração de *tokens* o “espaço” é sempre descartado, como pode ser visto na transformação acima.

Entretanto, é importante ressaltar que em algumas línguas não se utiliza o espaço como delimitador, como exemplo, japonês e o chinês. Neste caso, Indurkha e Damerau (2010) divide a tokenização em duas abordagens: uma para as línguas em que o espaço é o delimitador e outra para aquelas que não utilizam o espaço como delimitador.

Outro problema gerado pelos *tokens* é a dimensionalidade, uma vez que a divisão do texto em palavras, leva à criação de um grande número de dimensões para análise. Na subseções (2.6.3). serão apresentadas algumas técnicas de redução de dimensionalidade.

Por fim, o principal objetivo de criar tokens é a tradução de um texto em dimensões possíveis de se avaliar, analisar, para obtenção de um conjunto de dados estruturados (JACKSON; MOULINIER, 2007).

2.6.2 Remoção de *stopwords*

Em um *corpus* criado para minerar textos, devemos utilizar na atividade de Pré-Processamento, uma técnica conhecida na literatura como remoção de *stopwords*. Pois, ao manipular uma base

textual, encontra-se muitos *tokens* que não possuem valor para o contexto, sendo úteis apenas para a compreensão geral do texto.

Uma lista de *stopwords*, também conhecida como *stoplist* é constituída por palavras que adicionam pouco valor à análise. Normalmente, correspondem aos **artigos, preposições, pontuação, conjunções e pronomes** de uma língua (INDURKHYA; DAMERAU, 2010).

Segundo Wives e Loh (1998), uma *stopword* é considerada como “palavra vazia”, além de não colaborarem para a análise da polaridade de um texto, elas aparecem em praticamente todos os documentos, ou na maioria deles. São exemplos de *stopwords* em português “a”, “e”, “de”, “da”, dentre outras. A remoção dessas palavras traz um ganho de performance do sistema como um todo e reduz consideravelmente o tamanho final do léxico.

A criação de uma *stoplist* se dá através de tabelas de contigência que, depois, dão suporte para remoção das *stopwords* (MANNING et al., 2008). Geralmente, define-se a *stoplist* por um especialista no domínio da aplicação e, após essa definição, a remoção poderá ser realizada de forma automática, através da frequência de aparição das palavras no léxico. A Tabela 1 ilustra uma pequena *stoplist* definida manualmente e a identificação e descarte de *tokens*.

Tabela 1 – Identificação e remoção de *stopwords* (os *tokens* descartados estão sublinhados)

<i>StopList</i>	Texto
de, da, do, uma	[eu] [acho] [que] [tem] [<u>de</u>] [diminuir] [a] [maioridade] [é] [pra] [14]
para, um, tem	[Eu] [sou] [contra] [a] [redução] [<u>da</u>] [maioridade] [penal]
? ! : ;	
e, o, com	

Fonte: Autor

É importante ressaltar que, para aplicar a técnica de remoção das *stopwords*, deve-se analisar o que deseja manter do texto original, pois palavras importantes para a aplicação podem ser consideradas *stopwords*, ou dependendo do contexto, palavras que geralmente não compõem uma lista de *stopwords* podem ser adicionadas a ela.

Existem várias listas de *stopwors* disponíveis na internet, o que elimina a necessidade de construir uma lista manualmente, entretanto, para este trabalho construiremos uma que atenda o domínio da aplicação.

2.6.3 Redução do léxico

Conforme mencionado na seção(2.6.1), um dos problemas relacionado ao processamento de linguagem natural é o grande número de *tokens* que não possuem valor para análise. Pois, se considerarmos que cada *token* em um texto será mapeado para uma classe, gerando assim uma estrutura de dados de grande porte, que, por conseguinte, demandaria um elevado poder de processamento da máquina. Neste sentido, a redução de dimensionalidade torna-se muito

importante em processos de classificação automática, não somente para determinar os melhores atributos para modelagem, mas também para aspectos de escalabilidade dos modelos resultantes (KIM; STREET; MENCZER, 2000).

Yu e Liu (2004) descrevem que a quantidade excessiva de atributos causa lentidão no processo de treinamento, bem como na qualidade do conhecimento extraído. Dessa forma, a redução de atributos assume uma papel importante para o sucesso do processo, na medida em que os textos apresentam grande dimensionalidade e variabilidade de termos.

2.6.4 Normalização de palavras

Normalização é a etapa da redução do léxico que identifica e agrupa palavras que possuem relação entre elas.

Em geral, a aplicação das técnicas de normalização introduz uma melhora significativa nos sistemas de Mineração de Texto. Essa melhora varia de acordo com o escopo, o tamanho da massa textual e o que se pretende obter como saída do sistema (JUNIOR, 2007).

Segundo Manning et al. (2008) existem diversas técnicas para normalizar os dados e, essas técnicas vão de acordo com a necessidade da aplicação. Dentre as várias técnicas para realizar a normalização do dados, destacam-se os processos de **Stemming** e **Extração de Características**.

Stemming

Após a retirada das *stopwords*, pode-se realizar a técnica de stemming para reduzir cada palavra do léxico, originando assim os “termos”. A raiz de uma palavra é encontrada, na maioria das vezes, eliminando os prefixos, sufixos que indicam variação na forma da palavra, como plural e tempos verbais.

Em geral, por se tratar de um processo heurístico que simplesmente corta as extremidades das palavras na tentativa de alcançar o objetivo pretendido, os algoritmos utilizados nesta técnica, não se preocupam com o contexto no qual a palavra se encontra. Bem elaborado, o processo de *Stemming* traz benefícios no pré-processamento, sendo possível reduzir drasticamente o tamanho do léxico e também o esforço computacional, aumentando assim, a precisão dos resultados, exceto quando a retirada de prefixos e sufixos mudam a essência original da palavra.

A Tabela 2 exemplifica o processo de *stemming*, onde a segunda coluna apresenta o resultado da aplicação do algoritmo de stemming.

Analisando a tabela, percebe-se que nas frases 3 e 4 as palavras “ganhei” e “ganhando”, são convertidas para o mesmo radical “ganh”. Note que o mesmo ocorre com as palavras “perdi” e “perdemos” das frases 5 e 6, sendo atribuídas ao radical “perd”. Finalmente, as frases 2 e 9 transformam “dirigir” e “dirigindo” em “dirig”. Nestes três exemplos, todas as palavras possuem o mesmo sentido e a aplicação do algoritmo reduz consideravelmente a quantidade de palavras a serem processadas posteriormente.

Tabela 2 – Demonstração do algoritmo de *stemming*

ID	Frase Normalizada	<i>Stemming</i>
1	Ideia genial	Ide gen
2	belo dia dirigir	bel dia dirig
3	ganhei desconto carro	ganh descont carr
4	ganhando aposta	ganh apost
5	perdi novamente aposta	perd nov apost
6	perdemos jogo seremos eliminados	perd jog ser elimin
7	valor novo carro subiu	val nov carr sub
8	perdi novamente aposta	perd nov apost
9	chove perigoso dirigindo	chov perig dirig

Fonte: Autor

Apesar de o *Stemming* ser bastante útil na maioria dos casos, observe que nas frases 5 e 7 “novamente” e “novo” possuem sentido diferentes e, mesmo assim, foram transformadas no mesmo radical “nov”. Este fato demonstra a complexidade e nível de detalhamento que um sistema para analisar sentimentos em textos deve possuir. Uma possível solução seria adicionar a palavra “novamente” como *stopword*, pois ela não apresenta grande significância para o entendimento da frase 5.

2.6.5 Classificação automática de documentos

A classificação automática de textos reporta-se ao procedimento no qual um algoritmo classificador determina a qual classe um documento é pertencente. O principal objetivo da classificação é atribuir uma classe a um conjunto de documentos [Prabowo e Thelwall \(2009\)](#) Especificamente no caso deste trabalho, trata-se de distribuir automaticamente um conjunto de documentos entre classes.

A classificação pode ser dividida em um nível conhecido como aspectos, que trabalha em uma análise de maior granularidade dos documentos, quando a tarefa consiste em classificar cada característica do documento, e a classificação em nível de sentença ([MARTINS, 2003](#)).

Existem diversas estratégias para classificar um documento textual e, nessa dissertação, utilizaremos um classificador baseado em um modelo estatístico que trabalha com métodos indutivos, através de uma abordagem de aprendizado supervisionado, no qual um novo documento é classificado de acordo com as características aprendidas por este classificador, construído e treinado a partir de dados rotulados ([MARTINS, 2003](#)). O algoritmo em questão será o *Naïve Bayes*, que através dos dados de treinamento irá estimar a probabilidade de um documento pertencer a uma determinada classe.

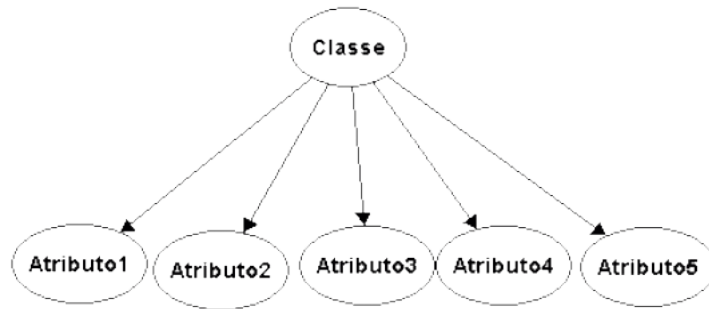
2.6.6 Naive Bayes

O classificador *Naive Bayes* é um gerador probabilístico para textos, sendo um dos mais utilizados em *Machine Learning*, devido a sua abordagem simplista com que trata as características dos modelos.

O termo *Naive* que, em português significa ingênuo, faz referência ao fato deste modelo assumir que existe uma independência condicional dos atributos, ou seja, a distribuição conjunta dos atributos é igual ao produto da distribuição de cada um deles (CHAKRABARTI, 2002).

Esse classificador é baseado no *Teorema de Bayes*, criado por Thomas Bayes no século XVIII, sendo este, considerado o mais eficiente na precisão e rotulação de novas amostras (CHAKRABARTI, 2002). Os classificadores *Naive Bayesianos* partem da hipótese que todos os atributos são independentes, dado a variável classe, e sua representação gráfica é exemplificada na figura 3.

Figura 3 – Estrutura do classificador *Naive Bayes* com 5 atributos e uma classe



Fonte: Chakrabarti (2002)

A distribuição conjunta de probabilidade do classificador *Naive Bayes* é dada por,

$$P(A_1, \dots, A_n, C) = P(C) \cdot \prod_{i=1}^N P(A_i|C)$$

No caso do classificador *bayesiano*, com atributos discretos e classe C, assumindo valores 0,1, a probabilidade de classificarmos um novo (caso), $A = a, \dots, A = a$ em $C=1$:

$$P(C = 1|A_1 = a_1, \dots, A_n = a_n) = \frac{P(C=1) \cdot P(A_1=a_1, \dots, A_n=a_n|C=1)}{P(A_1=a_1, \dots, A_n=a_n)}$$

E a probabilidade de classificarmos um novo (caso) em $C=0$ é,

$$P(C = 0|A_1 = a_1, \dots, A_n = a_n) = \frac{P(C=0) \cdot P(A_1=a_1, \dots, A_n=a_n|C=0)}{P(A_1=a_1, \dots, A_n=a_n)}$$

Com isso, uma nova observação (caso), $A = a, \dots, A = a$, é classificada na classe $C=1$ segundo o critério

$$\frac{P(C=1|A_1=a_1,\dots,A_n=a_n)}{P(C=0|A_1=a_1,\dots,A_n=a_n)} \geq 1$$

No caso do classificador bayesiano *Naive Bayes*, um novo (caso) $A = a, \dots, A = a$ é classificado em $C=1$ segundo o seguinte critério:

$$\frac{P(C=1)}{P(C=0)} \cdot \prod_{i=1}^N \frac{P(A_i=a_i|C=1)}{P(A_i=a_i|C=0)} \geq 1$$

2.6.7 Performance de classificadores

Avaliar a performance do classificador é muito importante na classificação de textos, pois, com as métricas é possível averiguar o quão este classificador é capaz de caracterizar um novo exemplo, quando lhe é apresentado (LIU, 2012). A maioria das métricas de avaliação utiliza-se de uma matriz contendo a quantidade de amostras classificadas corretamente e incorretamente, denominada de matriz de confusão. Essa matriz considera amostras positivas e negativas de uma das classes, ou seja, amostras positivas são pertencentes a uma das classes e amostras negativas são todas as outras pertencentes a outras classes. Desse modo, a matriz poderá ser construída para cada uma das classes do problema a ser avaliado.

Tabela 3 – Matriz de confusão

Classe correta	Positiva	Negativa
Positiva	Verdadeiras Positivas (TP)	Falsas Negativas (FN)
Negativa	Falsas Positivas (FP)	Verdadeiras Negativas (TN)

Fonte: Autor

Na matriz de confusão apresentada na tabela 3 TP representa o número de amostras positivas classificadas corretamente, FP é as amostras de outras classes classificadas na classe positiva, FN é a quantidade de amostra da classe positiva classificada em qualquer outra classe e TN é o número de amostras das outras classes classificadas corretamente.

A partir da matriz de confusão as métricas de precisão e acurácia, comumente utilizadas na avaliação de classificadores (LIU, 2012), podem ser definidas. Essa avaliação deverá ser realizada logo após a submissão do corpus ao treinamento, utilizando-se do resultado da classificação do conjunto de teste. Para tanto, existem diversas métricas que dão suporte nesta etapa, conforme listadas a seguir:

Precisão: Porcentagem de amostras positivas classificadas corretamente sobre o total de amostras classificadas como positivas. É definida pela fórmula:

$$\text{Precisão} = \frac{(Precision^+ \times T_p) + (Precision^- \times T_n)}{\text{total de tweets na base de dados}}$$

A medida Precisão é a média ponderada das medidas Precision+ e Precision-, em que TP e TN referem-se, respectivamente, à quantidade de tweets positivos e negativos na base de dados.

$$\text{em que Precision+} = \frac{TP}{TP + FP} \text{ e Precision-} = \frac{TN}{TN + FN}$$

O valor representado por TP (TN) refere-se à quantidade de *tweets* positivos (negativos) corretamente classificados e FP (FN) é a quantidade de *tweets* negativos (positivos) classificados como positivos (negativos), ou seja, é a quantidade de tweets classificados de forma errada.

Com a precisão, o analista poderá verificar o esforço executado para determinada busca. Isto porque, se 70% dos itens analisados e retornados forem relevantes, o analista teria, basicamente, desperdiçado 30% de seu esforço analisando itens irrelevantes. Logo, quanto maior a precisão, menor será o esforço realizado para analisar os itens. Deste modo, a precisão poderá auxiliar o analista na interação com o sistema ao passo de identificar o quanto ainda falta para detectar e filtrar os itens irrelevantes e retornar apenas os itens relevantes. A desvantagem dessa medida é que ela não leva em consideração as classes que deveriam ter sido reprovadas mas foram aprovadas.

Acurácia: Mede o quão efetivo o sistema é do ponto de vista da classificação geral, considerando o número de acertos sobre as amostras positivas e negativas de todas as classes. É definida pela fórmula:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}$$

Essa medida reflete um dos objetivos do trabalho, que visa, aproximar-se da capacidade humana de avaliar a subjetividade de um texto corretamente, variando entre 72% à 85% (GONÇALVES et al., 2006).

2.7 Trabalhos Relacionados

A *internet* é um meio onipresente, rápido, barato e anônimo para disseminação de quaisquer mensagem que um usuário deseja comunicar, e isso tem atraído muitos pesquisadores para a área de Mineração de Textos, sobretudo pela grande quantidade de informações em formato textual circulando no ambiente *online*, seja em fóruns, *blogs* ou redes sociais.

Para trabalhos relacionados buscou-se pesquisas sobre a Mineração de Textos, por duas razões. Em primeiro lugar, o discurso *web* é rico em conteúdo, emoção e opinião. Em segundo lugar, a análise desse tipo de texto é altamente relevante para pesquisas acadêmicas, uma vez que um texto opinativo desempenha um papel importante em influenciar as percepções das pessoas sobre a maneira em que comunicamos e na tomada de decisão. Além disso, buscou-se especificamente, pesquisas cujo o objetivo era prover técnicas para classificação de documentos.

Esse capítulo é organizado primeiramente por um quadro contendo as características principais sobre análise de sentimentos. Em seguida, um quadro com as principais características encontradas nos projetos relacionados a esta pesquisa e, na sequência uma descrição de cada um desses trabalhos.

A classificação de sentimento possui diversas características importantes, incluindo várias tarefas, técnicas e domínios de aplicações. Estas características encontra-se de forma resumida na taxonomia proposta por (ABBASI; CHEN; SALEM, 2008), apresentada na tabela 4.

Tabela 4 – Taxonomia de polaridade e classificação de sentimentos

Categoria	Característica	ID
Sintática	Word/POS tag n-grams.	C1
Semântica	Polaridade, grupos de avaliação, orientação semântica.	C2
Categoria	Técnicas	ID
Aprendizagem de Máquina	<i>SVM, Naive Bayes</i> , Vetores, <i>Stemming</i> , <i>Term Connection</i> e etc.	T1
Índice de similaridade	Contagem de frequência.	T2
Categoria	Domínios	ID
Discurso <i>web</i>	Fóruns e <i>blogs da web</i> .	D1
Avaliações	Comentários de produtos, filmes, músicas.	D2
Ontologias	Especificação formal e explícita.	D3

Fonte adaptada de: Abbasi, Chen e Salem (2008)

Com base na taxonomia proposta, a figura 6 mostra os estudos anteriores selecionados, que serviram de base para o desenvolvimento dessa dissertação. Discutimos os estudos relacionados em detalhes a seguir.

Figura 4 – Características dos trabalhos relacionados

Autores	Características		Técnicas		Domínios			Linguagens
	C1	C2	T1	T2	D1	D2	D3	
Abbasi et.al. 2008	✓	✓	✓		✓			Diversas
Ahmad, 2012.	✓	✓		✓			✓	Inglês
Efron, 2004.	✓		✓				✓	Inglês
Pak et. al. 2010	✓	✓		✓	✓			Inglês

Fonte: Autor

Observando a figura 4, percebemos que os trabalhos possuem algumas características em comum, no entanto, somente o trabalho de Abbasi, Chen e Salem (2008) utilizou das técnicas

de classificação de textos para diversos idiomas. A metodologia proposta por eles, foi avaliada através de um conjunto de dados sobre filmes norte-americanos e do Oriente Médio, por meio de postagens em fóruns *web* Árabe. Os resultados alcançados foram considerados satisfatórios, com uma precisão de mais de 91% de acertos.

O trabalho de [Ahmad e Doja \(2012\)](#) apresentou uma técnica para analisar a subjetividade de um texto, demonstrando que a detecção pode ocorrer em comentários extraídos de textos curtos e, ainda sim, manter a polaridade em um nível comparável ao da revisão completa. Para isso, o autor criou para o experimento uma base de dados contendo comentários sobre diversos modelos de câmeras digitais, sendo ela pré-rotulada entre 1000 frases subjetivas e 1000 objetivas e, criaram um software em Java para extrair as características do conjunto de dados, convertendo as frases em vetores numéricos. A avaliação da performance do sistema através da medida de *Recall* foi baixa, devido uma grande quantidade de palavras estarem escritas em uma linguagem não formal, com isso o analisador não conseguia identificar as sentenças corretamente.

O trabalho de [Efron \(2004\)](#) apresentou uma técnica para classificar documentos de hipertexto para estimar a orientação cultural, através de discursos polarizados sobre política. Usando um modelo probabilístico, dois experimentos foram relatados. O primeiro testou a capacidade do modelo discriminar entre apoiadores políticos de esquerda e os da direita. Neste modelo, foi testado 695 documentos *web* partidários, e 162 *weblogs* políticos. O classificador apresentado alcançou uma precisão acima de 90%. No segundo experimento, é proposto um modelo de classificação de páginas *web* de artistas musicais do gênero “alternativo”. Para este experimento, foi testado um conjunto de 515 páginas, alcançando uma precisão de 88%. No entanto, para melhorar os resultados alcançados, o autor sugere que seja abordado métodos lexicais e um corpus com maior número de dados.

E por último, [Pak e Paroubek \(2010\)](#) Apresentaram um método para treinar um classificador de sentimentos automaticamente, a partir de um *corpus* coletado do *microblogging* *Twitter*. O classificador criado foi capaz de determinar sentimentos positivos, negativos e neutros, sendo ele baseado no algoritmo de *Naive Bayes*, utilizando *N-gram* e *POS-tags* como características. Avaliações experimentais mostraram que as técnicas utilizadas foram eficientes. Além disso, o autor afirma que utilizando o algoritmo *Naive Bayes*, teve-se um melhor desempenho do que outros algoritmos já propostos.

Os trabalhos mencionados acima serão os pilares para o desenvolvimento desta dissertação, que possuirá grande semelhança com os mesmos. A nossa pesquisa tem a intenção de trabalhar de forma semelhante com o trabalho de [Abbasi, Chen e Salem \(2008\)](#), [Pak e Paroubek \(2010\)](#). Assim como desses autores, pode-se reutilizar nesta dissertação a criação de um modelo que tenha como base um *corpus* rotulado, o algoritmo *Naive Bayes* e o cálculo de acurácia para analisar a precisão da ferramenta desenvolvida. Além características encontradas nesses trabalhos.

3 Revisão sistemática de literatura sobre técnicas e métodos aplicados na análise de sentimentos

3.1 Introdução

O rápido crescimento das mídias sociais junto com a expansão da tecnologia de informação têm alavancado estudos sobre análise de sentimentos ou, em outras palavras, mineração de opinião, que tem por objetivo determinar o que as pessoas pensam sobre um determinado assunto. Sentimentos e opiniões estão contidos em conteúdos gerados por pessoas sobre serviços, produtos, política, educação, esportes e etc. Normalmente, as pessoas que frequentam sites de mídias sociais, estão compartilhando opiniões positivas e negativas a todo instante. Portanto, as características de um objeto tem um papel significativo na análise de sentimento. Esse artigo de revisão discute técnicas existentes e abordagens para ferramentas de análise de sentimentos. Adotou-se um processo de revisão sistemática da literatura para identificar as principais técnicas, domínios e métricas utilizadas pelos pesquisadores, buscando encontrar lacunas de investigação para trabalhos futuros.

3.1.1 Análise de Sentimentos

A análise de sentimentos é um subcampo da PNL que se baseia em abordagens da recuperação e informação, da lingüística computacional para identificar opiniões expressas nos textos (HAN; KAMBER; PEI, 2011). O principal objetivo da análise de sentimentos é o de identificar opiniões a respeito de determinado assunto, como um produto, serviço (LIU, 2012).

Sentimento é uma opinião ou uma avaliação de um indivíduo sobre algum aspecto ou objeto Liu (2012). Já a análise de sentimento ou mineração de opinião envolve o Processamento de Linguagem Natural, para tratar problemas de extração automática, classificação de textos e emoções expressas, principalmente, em textos online (ABBASI; CHEN; SALEM, 2008).

Segundo Gonçalves et al. (2013) a análise de sentimentos está substituindo a forma de analisar dados provenientes da *web* sobre entidades, como produtos e aliada a inteligência de negócios pode trazer grandes benefícios para as empresas. Em um contexto público, ela pode extrair informações sobre políticos e inferir sobre a reação do público sobre um novo projeto de governo (BARNES; LESCAULT, 2011).

Técnicas da análise de sentimentos incluem classificação de sentimentos, extração de características, redução da dimensionalidade, sumarização, entre outras (BERRY; KOGAN, 2010).

A classificação de sentimentos identifica a polaridade como positivo, negativo ou neutro. A extração de características separa os objetos através dos aspectos que estão sendo expressados em um texto e, finalmente, a sumarização agrega os resultados obtidos a partir das duas etapas anteriores.

Nessa revisão sistemática de literatura, discutiremos as técnicas existentes em forma intuitiva e, finalmente, as principais características nesta área são realçados. O texto dividi-se em seções. A seção II descreve a metodologia adotada para o processo de revisão; a seção III explora as técnicas e métodos utilizados na análise de sentimentos.

3.2 Planejamento

Para uma melhor compreensão das práticas recomendadas na literatura, aplicadas na classificação de sentimentos, propõe-se uma revisão sistemática de literatura, de acordo com o presente protocolo.

A revisão sistemática responde a uma pergunta claramente formulada utilizando métodos sistemáticos e explícitos para identificar, selecionar e avaliar criticamente pesquisas relevantes, e coletar e analisar dados de estudos incluídos na revisão (KITCHENHAM, 2004). Há razões para a execução de uma revisão sistemática da literatura, sendo algumas delas:

1. Resumir as evidências existentes sobre uma tecnologia.
2. Identificar lacunas nas pesquisas existentes e sugerir novos temas para debate.
3. Fornecer uma estrutura a fim de posicionar novas atividades futuras.

3.2.1 Questão de Pesquisa

Para o presente protocolo de revisão sistemática de literatura, propõe-se a seguinte questão de pesquisa:

- Quais técnicas, domínios e métricas de avaliação estão sendo aplicados na análise de sentimentos?

3.2.2 Amplitude da Pergunta

A. **Intervenção:** Mineração de textos.

B. **Controle:** Nenhum.

C. **Efeito:** Identificação das técnicas de PLN que estão sendo utilizadas no processo de mineração de textos.

D. **População:** Pesquisadores e projetos que explorem a mineração de texto com PLN, redes sociais, recuperação da informação, dentro outros.

E. **Resultados:** Pretende-se elaborar uma tabela contendo as principais técnicas, algoritmos de PLN aplicados à mineração de textos, o domínio no qual foram utilizadas.

F. **Aplicações:** A pesquisa servirá de base para elaboração deste estudo e a identificação de novas linhas de pesquisa em áreas como: recuperação da informação, mineração de textos, sistemas de recomendação de textos em mídias sociais.

3.2.3 Estratégia de Busca

Nessa pesquisa, são considerados trabalhos disponíveis na forma *online*. Os artigos devem estar escritos em inglês e devem relatar a aplicação de técnicas de PLN para mineração de textos. O inglês foi o idioma escolhido nesta pesquisa como fonte principal, pois a grande maioria dos estudos publicados na área estão disponibilizados neste idioma. Além disso, os periódicos e conferências mais relevantes também recebem estudos em inglês e eventualmente em outro idioma.

A pesquisa será realizada utilizando as seguintes fontes, informadas sem qualquer ordem de prioridade:

- *Emerald* (<<http://www.emeraldinsight.com/>>)
- *Science Direct* (<<http://www.sciencedirect.com/>>)
- *Ebsco Host* (<<https://www.ebscohost.com/>>)

Nas bases de dados, a pesquisa será realizada utilizando as seguintes palavras-chave, relacionadas ao estudo de classificação de sentimentos, a saber":

1. ("Opinion mining tools" OR "Sentiment analysis tools" OR "text mining tools")
2. AND ("techniques to sentiment analysis in texts")

A primeira sequência seleciona os sistemas de classificação e análise de sentimentos em textos, enquanto a segunda restringirá a busca aos métodos utilizados por parte dos pesquisadores.

3.2.4 Critérios de Seleção

Após a coleta inicial, será realizado uma triagem por pesquisa no título, resumo (*abstract*) e palavras-chave (*keywords*).

Serão considerados os textos publicados entre 2005 à 2015, em periódicos e conferências. Contudo, serão considerados apenas os trabalhos que a base fornecer acesso ao texto completo, gratuitamente ou para a instituição de ensino em que a pesquisa será realizada. Serão excluídos os

resumos, resenhas, teses e textos não publicados. Nessa etapa, será também realizado o processo de identificação da relevância do trabalho, caso apresente os termos de busca mas não estejam relacionados à pesquisa, estes serão excluídos.

3.2.5 Estratégias para Extração de Dados

Após a seleção dos artigos para a revisão sistemática, os dados serão tabulados em planilha, buscando identificar os algoritmos, domínios e métodos aplicados na classificação de sentimentos em textos. Desta forma, após a realização da revisão sistemática, será possível identificar as propostas mais comuns na literatura, observados os critérios de inclusão/exclusão.

3.3 Resultados

3.3.1 Realização

Para a execução da revisão sistemática, foi observado o protocolo estabelecido na Seção (3.2). As pesquisas foram executadas com as expressões de busca, utilizando também as parametrizações de pesquisa das bases de dados. Os resultados foram obtidos na ordem de classificação por relevância, de acordo com o próprio mecanismo de busca.

Através dos critérios de inclusão e exclusão foram retirados estudos que faziam apenas referências e citações ao tema, que não tratavam de uma técnica específica, ou cuja aplicação se dava em um idioma de estrutura diferente do inglês, como o chinês e o grego.

A execução da string de busca nas fontes selecionadas para o desenvolvimento dessa pesquisa retornou um total de 160 trabalhos distribuídos entre os anos de 2005 à 2015. O filtro aplicado através dos critérios de inclusão e exclusão, ocorreu na seguinte sequência de leitura: primeiramente utilizando o título dos trabalhos, em seguida o resumo e palavras-chave, as conclusões e por fim o texto completo, reduzindo o corpus inicial da pesquisa para 28 estudos.

A Tabela 5 sumariza a quantidade de itens obtidos com as consultas. Os refinamentos 1 e 2 correspondem, respectivamente, à exclusão baseada no tipo de documento, título, resumo e palavras-chave e quanto à relevância, após a leitura do texto completo.

Tabela 5 – Quantidade de itens obtidos pelas buscas e após aplicação dos critérios de inclusão/exclusão

Base de busca	Itens Retornados	Itens Retornados 1	Itens Retornados 2
Emerald	22	13	4
Science Direct	68	37	21
Ebsco Host	70	19	3

Fonte: Autor

Com base na leitura dos artigos, foi possível elencar as principais características de uma ferramenta para análise de sentimentos. Os quais, podem ser visto de forma resumida na tabelas 6 e 7. A terceira coluna, especifica os algoritmos utilizados. A quarta coluna especifica os artigos que usam as técnicas de AS para análise geral do texto (G) ou se resolvem especificamente problemas de classificação binária (positivo/negativo). A quinta coluna ilustra o escopo dos dados utilizados para avaliar os *algoritmos*, os dados podem ser resenhas de filmes, artigos de notícias, páginas *web*, *micro-blogs* e outros. A sexta coluna mostra as métricas utilizadas nos trabalhos para avaliar a performance dos algoritmos.

Quanto ao teor dos documentos, observa-se uma gama de propostas que variam de técnicas para classificação de textos, buscas em textos, extração de conhecimento em dados do tipo textual, representação do conteúdo de documentos e outros processos semânticos. Deve-se ressaltar que estes processos não necessariamente ocorrem de forma isolada, tendo sido encontradas evidências de experiências que combinam estas atividades, dependendo do resultado desejado.

3.3.2 Questão: Quais técnicas, Domínios e métricas de avaliação estão sendo aplicados na análise de sentimentos?

Por meio da análise das evidências encontradas na revisão sistemática, nos resultados e nas conclusões dos estudos, pôde-se elaborar as tabelas 6 e 7 de técnicas, algoritmos e domínios de dados usados em PLN e aplicados na mineração de textos para resolver questões de extração, representação, busca e classificação. No entanto, deve-se salientar que nem todos os trabalhos mostram de forma detalhada o uso da técnica, muitas vezes ocultando informações como a forma com a qual os dados textuais são estruturados ou como a técnica foi avaliada e escolhida para o estudo.

3.3.2.1 Técnicas para classificação de sentimentos

Técnicas para classificação de sentimentos em textos pode ser divididas em três abordagens, são elas [Bhaskar, Sruthi e Nedungadi \(2015\)](#) : abordagem híbrida, baseada em léxico e aprendizagem de máquina (ML). A abordagem híbrida combina ambas as abordagens e é muito comum com léxicos semânticos. A abordagem baseado no léxico, depende de uma conjunto de termos conhecidos e pré-compilados, ela ainda é dividida em outras duas vertentes, baseada em dicionário e baseada em um *corpus*, utilizando métodos estatísticos ou semânticos para encontrar o sentimento de polaridade. A abordagem de aprendizado de máquina utiliza, principalmente, os algoritmos de SVM.

Dentro da abordagem de ML existe uma divisão entre os métodos, a aprendizagem supervisionada e a aprendizagem não supervisionada. Os métodos supervisionados fazem uso de um grande conjunto de treino rotulados, normalmente, classificados em duas classes (positivos e negativos). Já os métodos não supervisionados são utilizados, principalmente, quando não é possível rotular

Tabela 6 – Sumário de artigos

Referências	Algoritmo	Polaridade	Dados	Métrica
Kechaou et al. (2013)	Markov, SVM	G	Site de notícias	Precision, Recall, F-Measure
Hammami, Guermazi e Hamadou (2008)	Decision tree, ID3, C4.5	G	Páginas web	Validação Cruzada, Bootstrapping
Barbosa, Sánchez-Alonso e Sicilia-Urban (2015)	Naive Bayes	POS/NEG	Comentários de hotéis no Trip Advisor	Accuracy
Zaghloul, Lee e Trimi (2009)	SVM	POS/NEG	Resumo de publicações	Accuracy
Valsamidis et al. (2013)	Naive Bayes	POS/NEG	Blog de agricultura	Não informado
K.M. et al. (2015)	SVM	POS/NEG	Ontologia	Accuracy
Kumar et al. (2015)	J-48, random Tree, ADT Tree, Breadth First, Naive Bayes, SVM	POS/NEG	Comentários de produtos	TP Rate, FP Rate, Precision
Jain e Kumar (2015)	SVM, Naive Bayes, Random Forest, Decision Tree	POS/NEG	Tweets sobre H1N1	F-measure, Precision, Recall
Kothari e Patel (2015)	SVM	G	Ontologia	Mean Reciprocal Rank
Kansal e Toshniwal (2014)	Baseado em regras	G	Ontologia	Accuracy
Hadano, Shimada e Endo (2011)	SVM	G	Comentários de games	Accuracy
Sanchez-Monzon, Putzke e Fischbach (2011)	Semantic, Latent Dirichlet Allocation (LDA)	POS/NEG	Comentários produtos McDonald's	Não informado
Weichselbraun, Gindl e Scharl (2014)	Lexicon-based, semantic	G	Comentários site Amazon	Não informado
Shahana e Omman (2015)	SVM	G	Revisão de filmes	Accuracy
Barawi e Seng (2013)	Baseado em regras	POS/NEG	Comentários em vídeos no Khan Academy	Não informado
Bafna e Toshniwal (2013)	Taxonomy-based, corpus-based	POS/NEG	Comentários no site Amazon	Accuracy, Precision
Bhaskar, Sruthi e Nedungadi (2015)	Multiclass SVM	POS/NEG	Dados em Audio - SemEval-2007	Accuracy
Barawi e Seng (2013)	Naive Bayes, SVM, rule-based	POS/NEG	Comentários sobre produtos Turísticos-Trip Advisor	Precision, Recall

Fonte: Autor

Tabela 7 – Sumário de artigos

Referências	Algoritmo	Polaridade	Dados	Método
Tamames e Lorenzo (2010)	SVM, Naive Bayes	POS/NEG	Ontologia	Precision, Recall, F-Measure
Landeghem et al. (2011)	Graph-Based approach, SVM	G	Ontologia	Precision, Recall
Zhao et al. (2015)	Maximum Entropy	G	Ontologia	Precision, Recall, FP Ratio
Moreno-Ortiz e Fernández-Cruz (2015)	Lexicon-Based,	POS/NEG	Notícias econômicas	Não informado
Yaakub, Li e Zhang (2013)	Context-based method, NLP	POS/NEG	Comentários de produtos	F-measure, Precision, Recall
Bach e Phuong (2015)	SVM, Naive Bayes	POS/NEG	Comentários no site da Amazon	Accuracy, Precision, Recall
Bilal et al. (2015)	KNN, Naive Bayes, Decision Tree	POS/NEG	Web-blog	F-measure, Precision, Recall
Preethi, Uma e kumar (2015)	SVM	POS/NEG	Tweets	Precision, Recall
Vidya, Fanany e Budi (2015)	SVM, Naive Bayes, Decision Tree	POS/NEG	Tweets sobre telefonia móvel	Net Brand Reputation (NBR), Precision, Recall, F-measure

Fonte: Autor

um conjunto de dados previamente. A classificação é realizada por meio de um léxico de palavras pré-selecionadas ou por conjunto de termos extraídos através das heurísticas de linguísticas ([KUMAR et al., 2015](#)).

Através da revisão sistemática, constatou-se que em mais de 82% dos estudos (23/28) utilizaram algoritmos de aprendizagem de máquina, como as máquinas de vetor de suporte (SVM) e o *Naive Bayes* (NB) em conjunto. Tanto o SVM quanto o *Naive Bayes* têm sido exaustivamente utilizado para analisar comentários de produtos [Barbosa, Sánchez-Alonso e Sicilia-Urban \(2015\)](#), [Kumar et al. \(2015\)](#), [Marrese-Taylor et al. \(2013\)](#), [Bach e Phuong \(2015\)](#), para revisões de filme [Shahana e Omman \(2015\)](#) e para análises de tweets [Vidya, Fanany e Budi \(2015\)](#), ([PREETHI; UMA; KUMAR, 2015](#)).

Os algoritmos de SVM e NB tornaram-se as técnicas mais utilizadas na classificação de sentimentos, ainda sim, encontrou-se nessa revisão sistemática a utilização de outros algoritmos que foram aplicados na classificação sentimento, são eles: *Markov* [Kechaou et al. \(2013\)](#), *LDA* [Sanchez-Monzon, Putzke e Fischbach \(2011\)](#), *Random Tree* [Kumar et al. \(2015\)](#) e *Decision Tree* ([BILAL et al., 2015](#)).

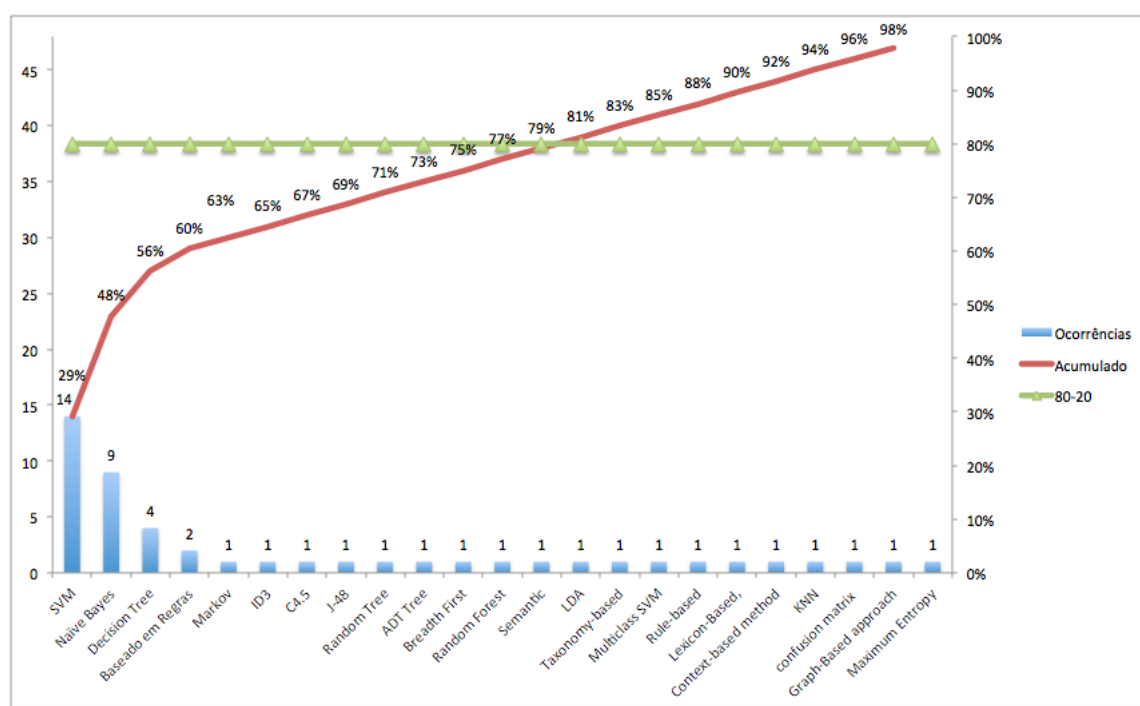
Métodos baseados em regras também são utilizados com outras características, tais como, a semântica. Normalmente, essas técnicas classificam uma frase com base na soma total entre as polaridades positivas e negativas [Hadano, Shimada e Endo \(2011\)](#), ([KANSAL; TOSHNIWAL, 2014](#)). Esse método exige que o texto seja previamente rotulado manualmente, frases etiquetadas

por polaridades (positivas ou negativas). Frases positivas são atribuídas a +1, enquanto frases negativas são atribuídas a -1. Consequentemente, todas as frases atribuídas com +1 são categorizadas com sentimento positivo, enquanto que as mensagens negativas é atribuídas à classe de sentimentos negativos (BAFNA; TOSHNIWAL, 2013).

A abordagem de orientação semântica Sanchez-Monzon, Putzke e Fischbach (2011) utiliza de métodos semelhantes para classificar as classes automaticamente, através das frases etiquetadas por polarização.

A figura 5 ilustra os algoritmos utilizados nos trabalhos estudados e a quantidade em que eles aparecem.

Figura 5 – Algoritmos mais utilizados



Fonte: Autor

3.3.2.2 Domínios

Estudos sobre análise de sentimentos tem sido aplicado em diversos domínios, incluindo comentários, revisões, notícias, discurso *web*, documentos e comentários sobre produtos Yaakub, Li e Zhang (2013), Weichselbraun, Gindl e Scharl (2014), Bafna e Toshniwal (2013), Bach e Phuong (2015). A análise de sentimentos tem sido também aplicada em fóruns *web*, *blogs* e recentemente, em mídias sociais como o *Twitter*. Neste caso, os estudos avaliaram os sentimentos sobre temas específicos, incluindo política, doenças, produtos (BILAL et al., 2015).

Jain e Kumar (2015) desenvolveram um método para detectar surtos de *Influenza-A (H1N1)* por meio da análise de dados gerados por usuários do *Twitter*, além de detectar através dos comentários a consciência da população indiana sobre o vírus.

Vidya, Fanany e Budi (2015) utilizaram os dados do *Twitter* para medir a reputação de marcas com base na satisfação dos clientes sobre alguns serviços específicos da telefonia: 3G, 4G e serviços de Internet, através da análise de sentimentos. Nesse estudo, foi utilizado três diferentes algoritmos: *Naive Bayes*, *SVM*, e um algoritmo baseado em árvore de decisão. Os resultados mostraram que *SVM* tem um desempenho melhor do que os outros dois classificadores (*Naive Bayes* e *Árvore de Decisão*), em processamento e precisão.

Contudo, a respeito de domínios, constatou-se através dos estudos que, o mais utilizado foi as ontologias. Ontologia é uma técnica de organização de dados que vem sendo muito utilizada nos últimos anos, principalmente no que diz respeito à representação formal de conhecimento K.M. et al. (2015), Kothari e Patel (2015), Kansal e Toshniwal (2014). Normalmente criadas por especialistas do domínio, tendo sua estrutura baseada na descrição de conceitos e dos relacionamentos semânticos entre eles, as ontologias geram uma especificação formal e explícita de uma conceitualização compartilhada (BHASKAR; SRUTHI; NEDUNGADI, 2015).

Na PLN, em um domínio ontológico as palavras são agrupadas e classificadas segundo uma ontologia de domínio, de modo que as sequências que tiverem o mesmo significado apresentem a mesma representação.

3.3.2.3 Métricas

Quanto aos métodos aplicados nos estudos para avaliar os resultados dos algoritmos, a uma predominância pelas métricas de *F-Measure*, *Precision*, *Recall* e *Accuracy*, como pode ser visto no gráfico 6. Nessa fase, é avaliado o resultado do processo de mineração de textos.

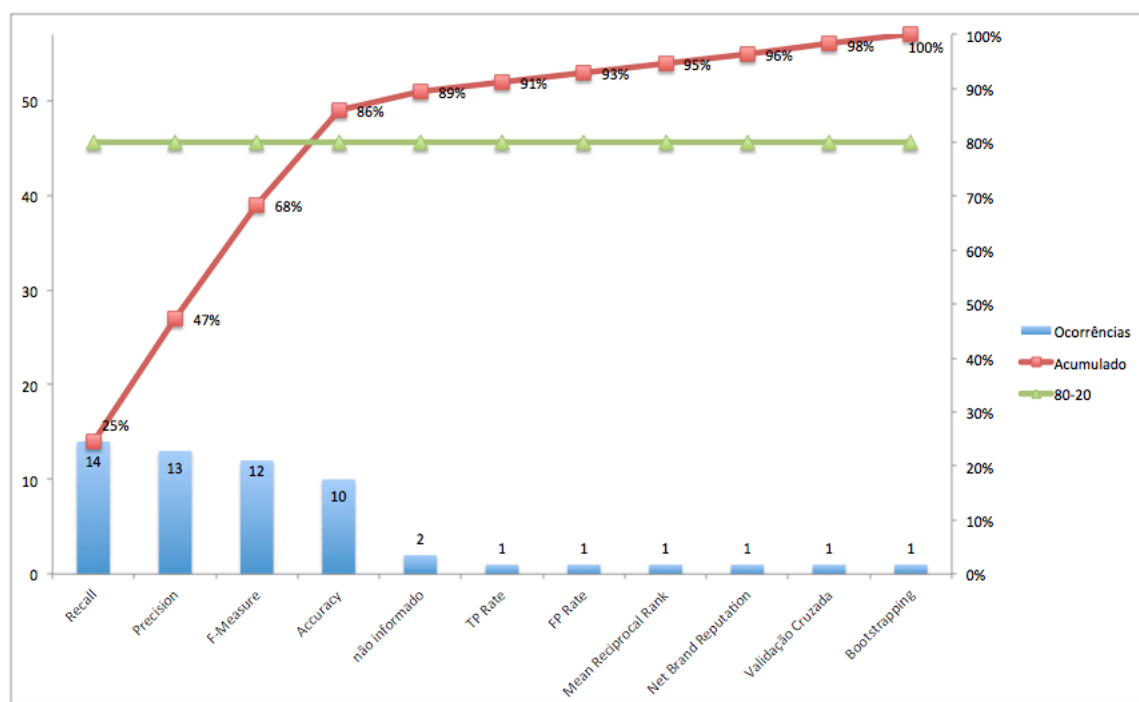
Para Witten e Frank (2011), métodos estatísticos podem ser utilizados como forma de avaliação dos algoritmos, isto é, saber se o processo funcionou ou não como previsto anteriormente. Nesse caso, as métricas podem informar para um usuário o percentual de classificações corretas para um determinado contexto.

Quanto aos resultados, foi observado que os trabalhos superaram a marca de 90% de precisão ao classificar um domínio, como em Tamames e Lorenzo (2010) e (BILAL et al., 2015). A maioria dos estudos descritos apresentam uma acurácia superior a 80%, o que pode ser visto como um resultado satisfatório.

3.4 Síntese do capítulo

A execução da *string* de busca nas fontes selecionadas para o desenvolvimento dessa pesquisa em 5 de Abril de 2015 retornou um total de 160 trabalhos distribuídos entre os anos de 2005 e 2015. O filtro realizado através dos critérios de inclusão e exclusão dos estudos, reduziu o corpus inicial da pesquisa para 28 estudos.

Figura 6 – Métricas de avaliação



Fonte: Autor

Os artigos selecionados para a revisão sistemática fazem uso das técnicas em Processamento de Linguagem Natural para atividades como a classificação de textos, buscas, representação e desenvolvimento de ferramentas.

Através das evidências encontradas no processo de busca e nos resultados, pôde-se elaborar uma lista de técnicas, algoritmos e métricas aplicadas na mineração de textos: *Naive Bayes*, *Decision Tree*, *Markov*, *Recall*, *Accuracy*, *Precision*.

Dos algoritmos, o *Naive Bayes* se mostrou o mais eficiente, primeiramente por ter sido aplicado em vários contextos ao longo de uma década e em segundo pela sua característica simples, sendo aplicável na maioria dos problemas relacionados à mineração de textos.

3.5 Considerações finais e trabalhos futuros

Esta revisão sistemática foi realizada tendo como base os 28 estudos selecionados através de critérios de inclusão e exclusão dentre 160 trabalhos que foram retornados pela string de busca.

Foram identificadas um total de 23 técnicas utilizadas para extração de padrões e conhecimentos em textos, bem como para a busca e classificação de termos. Dentre estas técnicas, os algoritmos de SVM e NB tornaram-se as técnicas mais utilizadas na classificação de sentimentos. Tais técnicas podem, inclusive, ser aplicadas para mineração de textos em sites de mídia sociais, num contexto em que predominam postagens com opiniões e expressões de sentimento.

Vale ressaltar, ainda, a variedade de pesquisas de mineração de textos na área médica, cujo o objetivo principal está voltado a extração automática de conhecimento em pesquisas empíricas e, os trabalhos que arquivam experimentos publicados através classificação e organização de base textuais.

Uma limitação presente nessa pesquisa foi o conteúdo disponibilizado pelos autores em seus artigos, nos quais as informações sobre técnicas e métodos foram ocultadas ou repassadas de forma incompleta, resultando assim em falta de dados e informações para elaborar uma discussão mais profunda em alguns casos, como na aplicação do algoritmo *Markov* no processamento de linguagem natural.

Como contribuição, este trabalho apresentou uma visão geral da aplicação de técnicas para a mineração de textos. Neste sentido, é apresentado ainda, uma lista descrevendo as principais técnicas identificadas, o contexto, a forma de aplicação e a estruturação de dados. Além disso, os achados dessa pesquisa pode servi como guia no processo de seleção e aplicação de técnicas para mineração de textos.

Como trabalhos futuros, prevê-se ainda, a execução da string em outras bases de estudos para identificar trabalhos que complementem os resultados encontrados até o momento. Em logo prazo, pretende-se impulsionar as pesquisas em mineração de textos, produzindo estudos combinados entre as técnicas aqui encontradas.

Por fim, outras discussões devem ser ampliadas para um melhor entendimento sobre a utilização das técnicas apresentadas neste trabalho, sobretudo aquelas que têm como foco a exploração automática de textos extraídos de mídias sociais.

4 Metodologia

No capítulo 2.1, introduziu-se o tema e apresentou-se o objetivo principal, que é implementar um modelo de sistema para classificar automaticamente sentimentos, em bases textuais, escrita em Português do Brasil, utilizando os conceitos da aprendizagem de máquina. Expôs-se, também, o problema de pesquisa (questão central da pesquisa): Quais os requisitos e os componentes de um sistema informatizado que seja capaz de analisar sentimentos em textos extraído de mídias sociais, utilizando as polaridades: Angustiado, Animado, Cansado, Entendiado, Encantado, Feliz, Irritado, Sonolento, Sereno, Satisfeito, Triste e Tenso? No presente Capítulo é explicitado os preceitos metodológicos que guiaram esta pesquisa.

4.1 Caracterização da Presente Pesquisa

A presente pesquisa segue o método *Design Science Research* (DSR) que [Wieringa \(2009\)](#), evangelizou, utilizando-se de uma estrutura aninhada do problema, que tem como objetivo a sua decomposição estrutural em problemas práticos e de conhecimento. O artefato escolhido são os modelos de mineração de textos para análise de sentimentos.

Quanto ao rigor e relevância da pesquisa, justificam-se os aspectos específicos dessa dissertação mediante as diretrizes proposta por [\(HEVNER, 2007\)](#).

4.2 Diretrizes de Hevner aplicadas a presente pesquisa

[Hevner \(2007\)](#) propõem uma estrutura composta por sete atividades em sequência nominal para as pesquisas em Sistemas de Informação. Nessa estrutura, o pesquisador avalia o problema observando aspectos, como pessoas, organização e tecnologias disponíveis. Assim, os artefatos são, posteriormente avaliados, sua importância é descrita e justificada.

Porém, para que seja factível (construção, justificativa e avaliação), é necessário consultar uma base de conhecimento já existente, através de fundamentos e métodos reconhecidos pela academia). Para apoiar as atividades de construção de artefatos bem como o apromiramento de teorias, [Dresch, Lacerda e Júnior \(2015\)](#) recomenda o método e a estrutura de [\(HEVNER, 2007\)](#).

As diretrizes proposta por [Hevner \(2007\)](#) são listadas na tabela 8. É realizado a associação de cada uma das diretrizes, seu detalhamento específico na presente pesquisa. O início se dá pela necessidade de gerar artefatos através da DSR com relevância para um problema específico e finaliza com a comunicação da pesquisa.

Tabela 8 – Diretrizes do Design Science Research

Diretrizes Hevner (2007)	Aplicação nessa pesquisa
1- Relevância do problema	Definição do problema específico de pesquisa e justificar o valor da solução
2- Objetivos da Solução	Inferir os objetivos da solução a partir da definição do problema
3- Artefato	Criar os artefatos da solução
4- Processo de Busca da Solução	Definição do problema específico de pesquisa e justificar o valor da solução
5 - Rigor da Pesquisa	Demonstrar a eficácia do artefato na solução do problema
6- Avaliação	Observar e mensurar quão bem o artefato suporta a solução do problema
7- Comunicação da Pesquisa	Comunicar o problema e sua importância

Fonte adaptada de: [Hevner \(2007\)](#)

4.3 Relevância do Problema

Um dos objetivos das pesquisas em Sistemas de Informação é gerar conhecimento que possibilite o desenvolvimento e implementação de ferramentas capazes de resolver problemas até então não resolvidos ou pouco explorado ([HEVNER, 2007](#)).

A imensa quantidade de dados circulando em formato textual na *web* nos últimos anos e o crescente uso dos mais variados sites de mídias sociais como o *Twitter*, torna humanamente impossível a extração de todas as informações relevantes contidas nos textos. Para facilitar o árduo trabalho de ler e associar os termos importantes, técnicas baseadas em mineração de texto têm auxiliado pesquisadores e, até mesmo instituições na busca de informações úteis.

Embora as técnicas de mineração de texto tenha sido aplicado com determinado sucesso ao estudo em diversas áreas, como análise de produtos e serviços [Giatsoglou et al. \(2017\)](#), [Hu e Liu \(2004\)](#), [Turney \(2002\)](#) ; em notícias [Godbole, Srinivasaiah e Skiena \(2007\)](#) e em áreas relacionadas à saúde [Goeuriot et al. \(2012\)](#) e política ([CHEN, 2001](#)). A inespecificidade das ferramentas desenvolvidas, principalmente, para língua inglesa torna difícil a correta extração de informações para o idioma Português.

Para compreender melhor sobre as técnicas e características das ferramentas para análise de sentimentos em textos, foi realizado um estudo comparativo entre *softwares* comerciais.

4.3.1 Comparativos entre as ferramentas comerciais de PLN

O estudo comparativo descrito na tabela 9, foi adaptado do trabalho de [Klemann, Reategui e Rapkiewicz \(2011\)](#), adicionando-se à este estudo a ferramenta AEMiner, desenvolvida nessa dissertação.

Tabela 9 – Análise comparativa de mineradores

Ferramenta	Online	Contagem de Termos	Apresentação de todos os termos	Apresentação de termos relevantes	Frequência dos termos	Contagem de Termos	Visualização gráfica dos termos	Visualização gráfica dos termos e relacionamentos
AEMiner	X	X			X	X	X	
TextAlyser	X	X			X			
IMiner	X	X	X		X			
TextSmart	X			X			X	
Sobek				X	X	X	X	X
BestChoice				X	X	X	X	X
uClassify	X				X	X		

Fonte adaptada de: [Klemann, Reategui e Rapkiewicz \(2011\)](#)

[Klemann, Reategui e Rapkiewicz \(2011\)](#) salienta a importância das ferramentas de mineração de textos possuírem uma camada *web* para interação com os usuários, considerando-se à tendência atual sobre computação em nuvem. A partir do estudo comparativo realizado por [Klemann, Reategui e Rapkiewicz \(2011\)](#), adicionamos as ferramentas *Bestchoice*, *uClassify* e AEMiner, desenvolvida nesse trabalho. Nesta seção, serão comentadas ferramentas comerciais para *text mining*. Uma descrição sucinta de cada uma delas é apresentada.

TextAlyser [Froelich e Ananyan \(2008\)](#) é uma ferramenta *online* para análise semântica de textos, com ela é possível identificar e determinar os termos mais importantes em um contexto através de uma contagem realizada pelo algoritmo. A ferramenta apresenta a frequência com que as palavras mais utilizadas ocorrem no texto, o número de palavras e sílabas. Diferentemente da ferramenta proposta nesse trabalho, o *TextAlyserI* não possui uma interface gráfica de fácil entendimento para o usuário e uma tabela para verificar a veracidade dos resultados.

IMiner [Wohl \(1998\)](#) é uma suíte da IBM composta por um conjunto de aplicativos que oferecem mecanismos de busca e análise de textos. Além disso, a suíte possui um *Web Crawler* embutido, o que torna fácil a busca de conteúdo na *web*. A ferramenta *Text Analysis Tools*, contida na suíte, realiza o processo de extração de característica, através de categorização e sumarização. É possível extrair palavras-chave em um documento e encontrar temas predominantes em uma coleção de dados. As primeiras versões dessa ferramenta, não possuía interface gráfica para o usuário, sendo necessário executar no ambiente *Windows* através de linha de comando em uma janela de *Prompt*.

A ferramenta *TextSmart* [Viechnicki \(1998\)](#) trabalha com dados no formato de perguntas e respostas, assimilando-se mais com uma enquete do que uma ferramenta de mineração de textos. Contudo, mesmo trabalhando num formato de pesquisa de opinião, a ferramenta tenta, na realidade, encontrar relacionamentos entre as respostas referentes a um mesmo tema. Para isso, ela executa o pré-processamento do texto, extraindo as *stopwords*, realiza o processo de *stemming*, apenas para o idioma inglês e categoriza automaticamente os termos em *clustering*. Como análise, a ferramenta gera uma lista com o ranqueamento geral dos termos, ou seja, o número de vezes que cada termo aparece em todas as respostas.

Já a ferramenta *Sobek* [Klemann, Reategui e Lorenzatti \(2009\)](#), não possui uma versão online. Para executá-lo é necessário um sistema operacional que rode em uma plataforma para *desktop*, podendo ser *Linux*, *Windows* ou *Mac*. Essa ferramenta é capaz de minerar textos nos formatos (*txt*, *doc* e *pdf*). Diferentemente das ferramentas apresentadas acima, essa se distigui por apresentar os conceitos encontrados na literatura, no que diz respeito a mineração de textos. Contudo, para utilizá-la, o usuário deve colar um texto na área de trabalho do *software* e executar o procedimento, não sendo possível utilizar uma base de dados, conforme é descrito na ferramenta desenvolvida neste trabalho. Além disso, ela também não exibe uma tabela descrevendo os resultados da análise.

A ferramenta *Bestchoice* [Silva \(2010\)](#) apresenta módulos semelhantes aos que foram desenvolvidos nesse trabalho. Carga, processamento e classificação. No entanto, para atribuir à polaridade as frases, esse *software* utiliza o *SentiWordNet*, que é uma base escrita em inglês para mineração de textos. Cada palavra opinativa foi traduzida para o inglês utilizando o *Google Translate* e assim era possível consultar na base do *SentiWordNet* os valores referentes à palavra: o *score* neutro (Obj), o *score* positivo (Pos) e o *score* negativo (Neg).

uClassify [Rodrigues et al. \(2016\)](#) é um classificador *online* que também determina a polaridade de um texto. Esse classificador possui uma base de dados em inglês, treinada com 2,8 milhões de documentos com dados do *Twitter* e *Amazon* sobre análises de produtos e críticas de filmes. Uma das principais características desse classificador, é realizar pesquisas, sondagens de marcas e ver as tendências em torno de campanhas de mercado. Porém, seu foco é na língua inglesa, o que dificulta sua utilização no Português, por exemplo. Há um projeto em andamento para construir uma *API* de tradução no *uClassify*, a idéia é oferecer um serviço acessível de tradução automática que pode rapidamente traduzir as solicitações para o idioma classificador. Ainda sim, este classificador pode ter seu uso limitado na língua portuguesa, devido às particulares dessa língua.

As ferramentas descritas nesta parte da tese representam algumas das mais conhecidas ferramentas comerciais existentes no mercado. O objetivo desta seção foi dar uma noção da variedade de abordagens existentes nas aplicações sobre mineração de textos.

Ferramentas para mineração de texto semelhantes ao trabalho aqui proposto, descritas na Seção [4.3.1](#) se diferem, na maioria, na forma de analisar a ocorrência dos termos baseando-se em

um dicionário de dados em linguagem semi-estruturado, o que torna a análise muito superficial, superestimando a predição de toda correlação dos termos. Além disso, poucas executam o processo de mineração de textos diretamente na plataforma *web*, o que dificulta análises em tempo real em larga escala, bem como a integração direta com sites de mídias sociais.

Outra característica desses estudos é que as técnicas aplicadas são avaliadas ou comparadas considerando base de treinamento balanceadas. Porém, como descrito por [Li, Lee e Huang \(2013\)](#) os *corpus* de opiniões presentes na *web* são desbalanceados, ou seja, *corpus* que apresentam ruídos e uma significativa diferença entre a quantidade de dados pertencentes a determinadas classes. Por fim, as ferramentas analisadas não procuram descrever quais as características que uma ferramenta precisa ter para avaliar um texto com mais de duas polaridades. Entende-se que os problemas aqui descritos são relevantes e que essas características podem influenciar consideravelmente na aplicação das técnicas e na identificação de padrões em textos e, essas lacunas constitui-se da primeira atividade da DSR, a qual buscou-se encontrar um problema e motivação para essa dissertação.

Propomos então, um método integrado na ferramenta desenvolvida nesse trabalho, denominada-se AEMiner, para classificar automaticamente textos extraídos de sites de mídias sociais. A ferramenta é capaz de identificar e correlacionar os termos mais representativos em uma frase, intercalando-os para aumentar as chances deles estarem de fato relacionados ao mesmo evento. Além disso, a ferramenta possui uma camada *web* que pode ou não está ligada diretamente em um site de mídia social, facilitando assim o processo de extração e montagem do *corpus* para classificação. Com este propósito, foram estabelecidos os objetivos do trabalho, conforme apresentados na Seção 1.2.

4.4 Artefatos

Nas pesquisas em Sistemas de Informação, desenvolve-se um artefato visando resolver um problema relevante. Na presente pesquisa, foi implementado um modelo de sistema que classifica textos escritos em Português do Brasil, através de polaridades.

A terceira etapa da estrutura proposta por [Hevner \(2007\)](#) constituiu-se basicamente da pesquisa sobre as características e da elaboração dos requisitos necessários ao sistema, partindo de uma proposta metodológica que se adequasse, principalmente, a questão de pesquisa desse trabalho, descrita no capítulo 2.1. Nessa etapa, alcança-se a produção dos artefatos planejados na DSR.

Segundo [Peffer et al. \(2006\)](#), os **artefatos** podem ser dos tipos: construtos (entidades e relações), modelos (abstrações e representações), métodos (*algoritmos* e práticas) e instanciações (implementação de sistemas e prototipação). O presente trabalho, segundo esta classificação, têm como principal artefato a elaboração **métodos**. Para a construção do artefato, foi realizada uma revisão sistemática de literatura, conforme protocolo registrado em 3.2. Esta revisão subsidiou as informações necessárias à consecução do objetivo deste trabalho.

4.5 Processo de Busca da Solução

Para [Hevner \(2007\)](#) o ciclo de *design* é o coração de todo projeto de pesquisa em Ciência de *Design* e também onde o trabalho é mais oneroso e intenso. Deve-se descrever todo o processo de condução da pesquisa, bem como métodos para a criação e avaliação do artefato.

O desenvolvimento da pesquisa foi dividido em ciclos de *design*. O primeiro ciclo de desenvolvimento é o ciclo intitulado de "Implementação" e está detalhado na seção [4.5.1](#). O segundo ciclo é denominado "Arquitetura do Sistema" sendo é apresentado na seção [4.5.2](#).

4.5.1 Ciclo 1 do Design: Implementação

O primeiro passo para a construção dos artefatos para classificar os sentimentos propostos na questão de pesquisa, (seção [1](#)), foi necessário um estudo sobre ferramentas, (seção [4.3.1](#)) já desenvolvidas para este fim. O entendimento das limitações proporcionou o *design* de artefatos adaptados à necessidade identificada.

Nas seções [4.5.1.1](#), [4.5.1.2](#), [4.5.1.3](#), [4.5.1.4](#), [4.5.1.5](#), [4.5.1.6](#) são apresentados conceitos e técnicas utilizadas no desenvolvimento dessa dissertação.

4.5.1.1 Aprendizagem automática

Grande parte das pesquisas realizadas em mineração de textos, são sobre sistemas de aprendizagem supervisionadas, isto é, a partir de um corpus de treinamento, um algoritmo sintetiza um determinado conhecimento de forma probabilística [Mitchell \(1997\)](#). Esse conhecimento passa para um programa extrator do tipo de informação selecionada para o aprendizado. Assim, o sistema terá autonomia para processar, em novos textos, as características desejadas e, atribui-las à uma determinada polaridade.

A figura [7](#) ilustra, em alto nível, as etapas utilizadas nos casos de aprendizagem supervisionada. O conhecimento probabilístico descrito nessa figura, no caso do pré-processamento, é armazenado em um léxico (acervo de palavras).

4.5.1.2 Léxico computacional

Um léxico computacional é composto de um repositório de palavras que busca armazenar todas as palavras (lexemas) existentes em uma língua. Lexemas são unidades mínima distintiva do sistema semântico de uma língua que reúne todas as flexões de uma mesma palavra [Longhi et al. \(2010\)](#). Obviamente, em uma dada língua existe infinitas palavras, portanto, o léxico computacional armazenará apenas o número de palavras diferentes no texto processado (*corpus*).

Além disso, o léxico computacional cria uma associação de cada palavra com outras informações pertinentes à aplicação, são os casos das polaridades encontradas em grande parte dos trabalhos sobre mineração de textos, como exemplo: positivo, negativo, neutro entre outras. Nesse sentido,

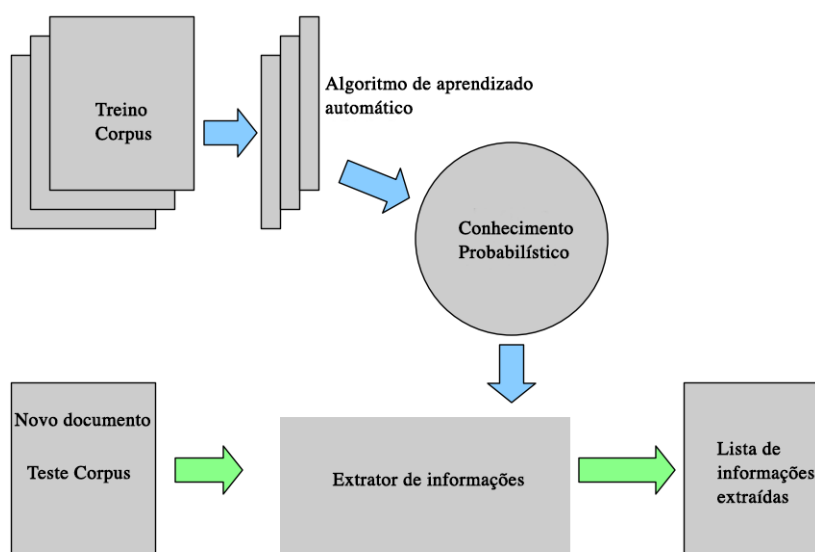


Figura 7 – Arquitetura em alto nível de um sistema de extração

Fonte: Autor

o léxico possui um papel importante de apoio à percepção da palavra. Mesmo sendo apoiado no léxico mental, que é como nosso cérebro armazena as palavras de uma língua, o léxico computacional busca atingir o mesmo objetivo [Aranha e Vellasco \(2007\)](#). Contudo, ele já leva consigo o nome ("computacional") por não ter nenhuma pretensão de ser fiel ao léxico mental [Gomes, Neto e Henriques \(2013\)](#). Isso quer dizer que o direcionamento acerca da modelagem dos dados desse trabalho não é necessariamente um problema cognitivo da mente humana.

A respeito da modelagem lexical, um sistema de processamento de linguagem natural, seja ele voltado para o comércio ou apenas de cunho acadêmico, requer, forçosamente, uma base de dados lexicais. No entanto, dada a dificuldade em manipular esses sistemas, no que diz respeito ao processamento do texto e a grande quantidade de informações relativas às palavras que são armazenadas e manipuladas pelo sistema, as exigências quanto à estruturação das bases de dados têm sido cada vez mais acentuadas. De modo geral, essa dificuldade imposta na estruturação se justifica pelo fato de não buscar a extração de informações encontradas nos dicionários de palavras, isto é, classes gramaticais e os significados das palavras que lá se encontram. O que se busca, é uma estrutura lexical ricamente elaborada, bem como a redução do léxico, sem perda de consistência do conjunto de dados ([ARANHA; VELLASCO, 2007](#)).

4.5.1.3 Linguagem *Python*

A linguagem de programação *Python* foi mostrada ao mundo no final dos anos 80. A ideia original surgiu no Centro de Matemática e Ciência da Computação em Amsterdã, Holanda. Atualmente, *Python* é usada para diversas tarefas, desde scripts simples a grandes servidores *web* escaláveis que provêm serviços ininterrupto 24x7. É usado também em interfaces gráficas de

usuários, em programação para *web* tanto no lado cliente quanto servidor e, ainda, por cientistas que escrevem programas para supercomputadores e até por crianças aprendendo a programar (BURIOL; ARGENTA, 2009).

Python possui uma ampla biblioteca, com mais de 100 módulos em constante evolução, sendo isso um dos principais motivos para sua popularidade. Alguns desses módulos incluem funções para cálculos matemáticos, expressões regulares, interfaces, threads, protocolos de redes, processamento *XML*, *HTML* e um kit para interface gráfica de usuário (tcl/tk) Buriol e Argenta (2009). Além disso, há possibilidade de utilizar módulos e pacotes de terceiros que, em sua maioria, são de código fonte aberto. Dentre esses, pode-se encontrar empacotadores para *Fortran*, interface com banco de dados *Oracle*, *MySQL*, *frameworks* para *web* e outros.

A linguagem *Python* foi escolhida para o desenvolvimento da ferramenta nessa pesquisa por se tratar de uma linguagem eficiente e de fácil manutenção na programação para uma amplitude de tarefas. Dentre as facilidades oferecidas pela linguagem *Python* está a manipulação de arquivos de texto, motivo principal da escolha desta linguagem.

4.5.1.4 *Natural Language Toolkit*

O *Natural Language Toolkit* é um kit de ferramentas e módulos para apoiar à investigação em linguística computacional e linguagem natural. NLTK é escrito em *Python* e distribuído sob a *GPL Open Source* Licença.

Loper e Bird (2002) descrevem o NLTK como uma suíte de aplicativos e módulos de código aberto, para prover o aprendizado da linguagem natural. A suíte foi construída para conceber: documentação, simplicidade, facilidade de uso, consistência, extensibilidade e, por esses motivos, é a mais utilizada no mundo acadêmico na área de PLN.

Além disso, a suíte foi projetada para disponibilizar coleções de módulos independentes, em que cada um define uma estrutura de dados específica, sendo os principais módulos: *parser*, que especifica uma interface representando os textos através de árvores; o rotulador, que é responsável por expandir as características e informações de um determinado token com dados adicionais; o classificador, que utiliza-se de uma interface para a classificação de textos em categorias, sendo realizado através do algoritmo *Naive Bayes*.

A suíte NLTK é ideal para pesquisadores em PNL, ou áreas afins. Ela foi usada com sucesso como plataforma de desenvolvimento para sistemas de prototipagem e para pesquisas em análises de sentimentos (LIDDY; MCCracken, 2005).

Para este trabalho, escolheu-se esta biblioteca pela sua curva de aprendizagem, sua sintaxe transparente e na facilidade de manipular funções através da linguagem de programação *Python*. Os códigos criados em *Python* podem ser encapsulados e reutilizados com facilidade, além da ampla biblioteca, incluindo ferramentas para programação visual e processamento numérico (BEAZLEY, 2006).

4.5.1.5 Stemmer

O algoritmo implementado na ferramenta para o processo de *stemming* de dados em Português [Orengo e Huyck \(2001\)](#) é o *Stemmer Portuguese*, mais conhecido pela sigla RSLP (Removedor de Sufixos para a Língua Portuguesa). Escrito originalmente em C, este algoritmo é composto de oito passos que devem ser executados seguindo uma ordem definida. O fluxograma mostrado na figura 8 apresenta a sequência que estes passos devem obedecer.

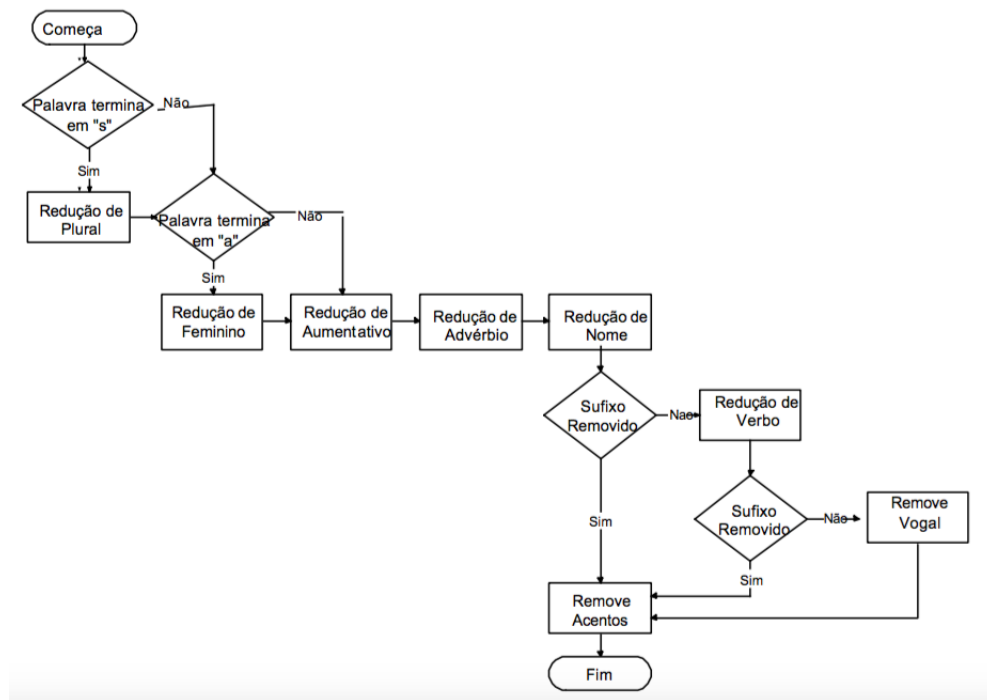


Figura 8 – Sequência de passos para o algoritmo de stemming RSLP

Fonte: Autor

No fluxo, é observado diversas regras proposta pelo *Stemmer*. Porém, nem todas são aplicadas dependendo do contexto. A cada etapa as regras são conferidas e apenas um valor é aplicado. Até o desenvolvimento desse trabalho, esse *stemmer* para o idioma Português possuía 199 regras, que estão apresentadas no Apêndice B.

Basicamente, cada regra mostra:

1. (O sufixo a ser removido;)
2. (O comprimento mínimo do *stem*;))
3. (Um sufixo de reposição a ser anexado ao *stem*, se aplicável;)
4. (Uma lista de exceções: para aproximadamente, todas as regras definidas;)

- O comprimento mínimo do *stem* : Serve para evitar que um sufixo seja removido quando um *stem* é muito curto. Antes da remoção do sufixo, é verificado os valores definidos pela observação das palavras terminadas com o dado sufixo. Tal medida, apesar de não ter suporte linguístico para esse procedimento, reduz erros de overstemming.
- Uma lista de exceções: para 98% das regras existentes, existirão exceções. Então, uma lista de exceções é adicionada para cada regra. Tais listas foram construídas através de um vocabulário de 32.000 palavras em português, disponível em [Orengo e Huyck \(2001\)](#). Os testes com o stemmer têm mostrado que a lista de exceções reduz erros *overstemming* em 5%. Um exemplo de uma regra foi explicando na seção 2.6.3 na tabela 2.

4.5.1.6 Algoritmo de *Stemmer Portuguese*

Descrição do Passo 1 – Redução de Plural

Como no Português a única palavra que denota plural mas não termina em (s) é a palavra (quaisquer), fica fácil para o algoritmo de *stemmer* remover todas as palavras que terminam em (s). Esse passo consiste em remover o (s) final das palavras. Contudo, algumas palavras que terminam em -s não são plurais, um exemplo, lápis. Com isso, faz-se necessário utilizar-se de uma lista de exceções.

Descrição do Passo 2 - Redução de Feminino

Todos os substantivos e adjetivos em Português têm um gênero. Nessa etapa, o foco do algoritmo é transformar formas femininas em seus respectivos correspondentes masculinos. Apenas palavras terminando em (-a) são testadas nesta etapa.

Descrição do Passo 3 - Redução de Advérbio

Essa é a etapa mais curta de todo o processo, pois, existe apenas um sufixo que denota advérbios (-mente). Novamente, a lista de exceções é utilizada, já que nem todas as palavras que terminam com (-mente) são advérbios.

Descrição do Passo 4 - Redução de Aumentativo/ Diminutivo

No Português, nomes e adjetivos apresentam muito mais variações que o equivalente em Inglês. As palavras possuem formas aumentativas, diminutivas e superlativas, por exemplo, casinha, onde (-inha) é o sufixo que indica um diminutivo. Esses casos são tratados nessa etapa.

De acordo com [Cunha, Cintra et al. \(1985\)](#) há 38 desses sufixos, contudo, alguns deles são obsoletos. Para evitar um overstemming nessa etapa, o algoritmo de stemming utiliza apenas os mais comuns que ainda são de uso comum.

Descrição do Passo 5 - Redução de Sufixo de Nome

Nessa etapa o algoritmo faz um teste com as palavras em relação à 61 substantivos e adjetivos. Se um sufixo é removido aqui, as etapas 6 e 7 não são executadas.

Descrição do Passo 6 – Redução de Sufixo de Verbo

No idioma Português existem diversas formas verbais, ao contrário do inglês que tem apenas quatro variações em verbos regulares, por exemplo (*talk, talks, taked, talking*). No Português os verbos regulares apresentam aproximadamente 50 formas diferentes [Cunha, Cintra et al. \(1985\)](#). Os verbos variam de acordo com o tempo, pessoa, número e modo e, as estruturas verbais são apresentadas como raiz+vogal temática+tempo+pessoa, por exemplo, and+a+ra+m.

Descrição da etapa 7 - Remoção de Vogal

Nessa etapa é realizado a remoção da última vogal (“a”, “e”, “o”) das palavras que não foram alteradas pelo stemming na etapa 5 e 6. Por exemplo, a palavra menino não seria modificada pelo stemming nos passos anteriores, portanto essa etapa removerá seu (-o) final, de forma que ela possa ser reunida às outras variações, tais como menina, meninice, meninão, menininho, que também serão convertidas ao stem *menin*.

Descrição da etapa 8 - Remoção de Acentos

Remover os acentos no processo de mineração de textos é muito importante e necessário, pois, existem casos em que algumas palavras possuem variações em acentuação e outras não, como em psicólogo e psicologia. Após essa etapa ambas variações seriam unidas para *psicolog*.

4.5.2 Ciclo 2 do Design: Arquitetura Geral do Sistema

O modelo de classificação automática e análise de sentimentos está de acordo com o Diagrama Hierárquico de Funções do sistema [9](#). Ele foi desenvolvido utilizando a linguagem de programação *Python* com auxílio na biblioteca NLTK para processamento de linguagem natural.

4.5.3 Coleta de dados

O sistema aceita duas formas de entrada de dados, a primeira opção é coletar automaticamente através de um *web crawler* os dados desejáveis, o outro formato pode ser inserido diretamente por um usuário através de uma base de dados já definida. Para que um usuário insira dados, é necessário informar em qual pasta do disco local contém os documentos (sejam estes coletados de uma mídia social ou não) que serão incorporados ao sistema. Para que um novo *corpus* seja formado, o usuário deve acessar o sistema através do menu Nova Base->Carregar *Corpus*.

4.5.4 Pré-Processamento

O módulo de Pré-processamento constitui-se de algumas das principais tarefas existentes em todo o processo de mineração de textos, as quais foram explicadas na na seção [2.6](#) dessa dissertação. A figura [10](#) ilustra as classes introduzidas ao diagrama principal [9](#) para executar as etapas do pré-processamento.

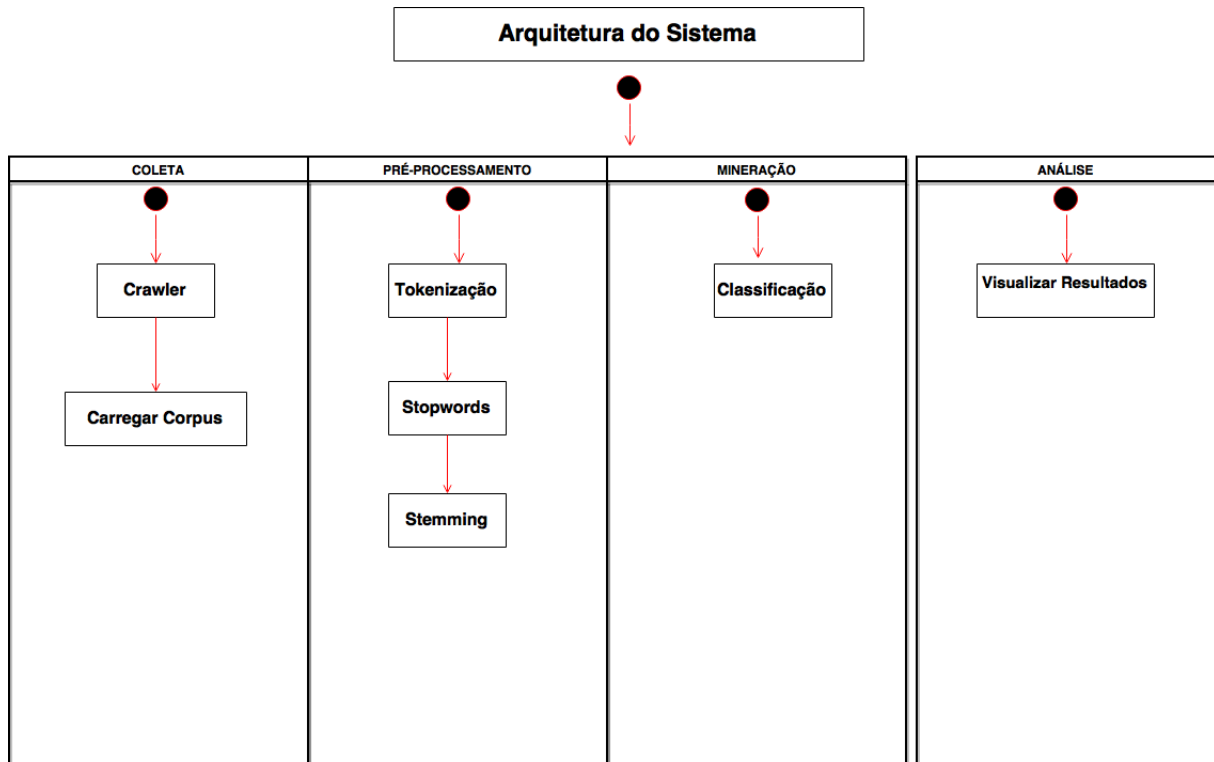


Figura 9 – Diagrama de funções do sistema implementado

Fonte: Autor

Nesse trabalho, são aplicadas três técnicas comumente utilizadas neste campo de pesquisa, ou seja, a remoção das *stop words*, a aplicação do algoritmo de *stemming* e a distribuição da frequência das palavras. Esta interface é definida no diagrama da figura 10 como IExtrator. Três classes implementam essa interface: *Stoplist*, *Stemmer* e *Tokenizador*.

A primeira constitui-se de um algoritmo que normaliza os textos de entrada. Tal procedimento consiste na remoção dos termos considerados irrelevantes, os quais são chamados de *stop words* e da remoção da pontuação através de uma expressão regular.

Através de um processo executado por meio de uma estrutura de repetição que varre a base de dados, é armazenado em uma variável a lista com as palavras individuais divididas e normalizadas para minúsculas. Em seguida a função `get_palavras` cria uma lista vazia, a qual irá receber cada uma das frases separadas por palavra com o respectivo sentimento associado. O objetivo desse laço de repetição é quebrar toda a frase e transformar cada palavra dela em um item da lista.

A segunda consiste em um algoritmo simples que divide o texto em *tokens*, reduzindo também a palavra ao seu radical.

A terceira, *DistFrequencia* é uma classe que implementa uma função que tem como objetivo pegar a lista criada no passo anterior e, através de uma variável, armazenar um dicionário reconhecido como *FreqDist*, em que as palavras são as chaves e o número de vezes em que elas foram repetidas (frequência) é o valor da chave. Esses valores são armazenados em uma outra

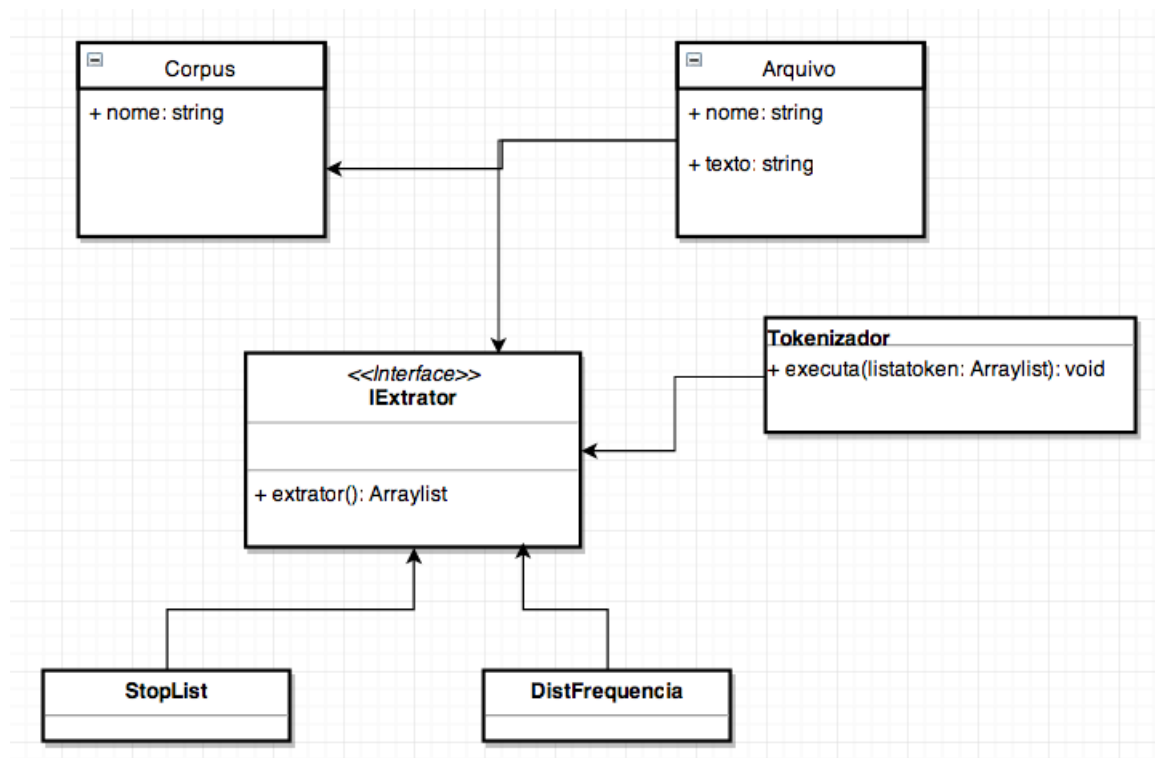


Figura 10 – Diagrama de classes da etapa de pré-processamento.

Fonte: Autor

variável com base na frequência de palavras, para formar uma nova lista sem repetições e com os radicais únicos existentes em todas as frases.

4.5.5 Mineração

Na camada de mineração contém a implementação do algoritmo *Naive Bayes*, explicado na seção 2.6.6 desta dissertação. A implementação do algoritmo introduziu uma função ao diagrama de classes da ferramenta, visto parcialmente na figura 11.

O algoritmo de *Bayes* classifica itens segundo a presença ou ausência de características no conjunto de dados. No caso dessa ferramenta cada característica é uma dada palavra (limpa) de idioma natural.

O classificador foi implementado utilizando a função *NaiveBayesClassifier* da biblioteca NLTK. A função deve receber a lista de palavras passada por parâmetro e, em seguida, deve-se fazer uma chamada ao método *nlk.classify.apply_features*. A função “*apply_features*”, recebe como parâmetro a lista das frases originais já pré-processadas, iniciando assim à fase de treinamento do algoritmo.

Logo após a fase de treinamento, o classificador já está apto à classificar documentos de forma automática. Com o classificador definido, o último passo é realizar os testes com novas frases e avaliar a previsão do sentimento que o classificador irá retornar.

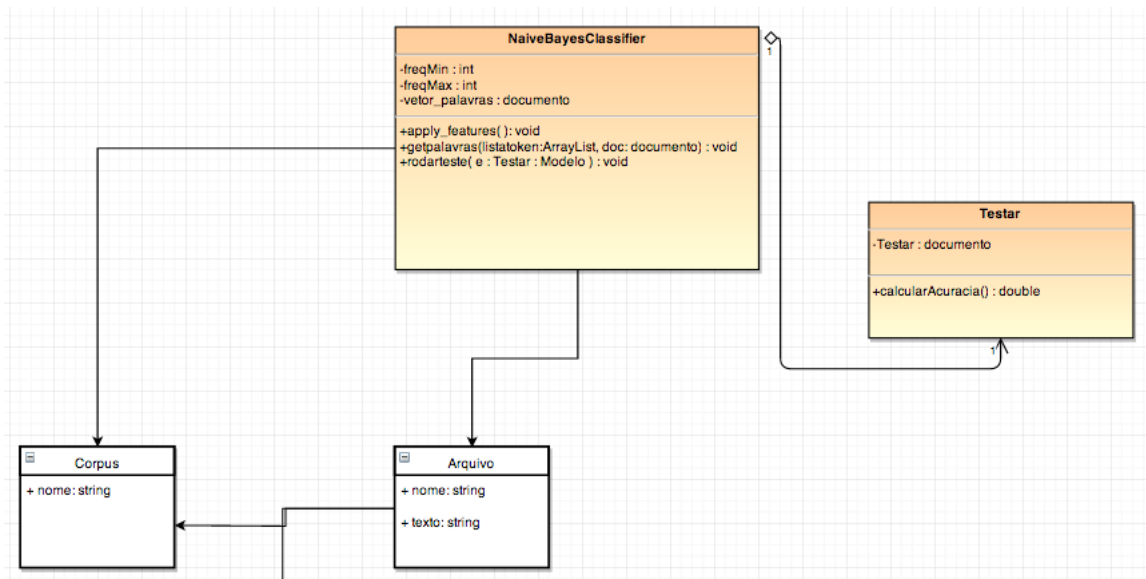


Figura 11 – Diagrama de classes parcial do sistema com a inclusão das classes que compoem o módulo de mineração.

Fonte: Autor

Por fim, a classe **Testar** contém os métodos para realizar testes que indicam a capacidade de classificar da ferramenta. Para tanto, um objeto dessa classe deve receber dados provenientes do classificador, conforme descrito no parágrafo anterior. Nesta chamada, o teste deve ser enviado como parâmetro. O resultado do teste fica contido no próprio objeto, sendo acessado através de métodos que realizam cálculos segundo determinadas métricas, como exemplo, Acurácia.

4.5.6 Visualização e análise dos resultados

Para a visualização e análise dos resultados, foi implementado um web site para auxiliar o usuário quanto à novos testes e a compreender melhor os resultados obtidos acerca do classificador e, sobretudo, o poder discriminatório do mesmo. Além disso, é possível verificar através de uma matriz de confusão [12](#), o número de classificações corretas em oposição às classificações preditas para cada classe.

Através da matriz de confusão, é possível deduzir sobre a performance do classificador. Como exemplo, a acurácia seção [2.6.7](#), que é o quanto ele classifica de forma correta os exemplos apresentados e, verifica através da soma das corretas classificações (quadrante A e D) dividido pelo total de exemplos apresentados (soma de todos os quadrantes, isto é, A, B, C e D).

4.5.7 Interfaces e funcionalidades da ferramenta

O Apêndice A traz as interfaces da ferramenta AEMiner e algumas notas explicativas referentes às mesmas. As funcionalidades implementadas na ferramenta e que estão disponíveis para o usuário são:

	Documentos classificados como sendo da classe X	Documentos classificados como sendo da classe Y
Documentos da Classe X	A	B
Documentos da Classe Y	C	D

Figura 12 – Matriz de confusão

Fonte: Autor

1. (Criar um conta de novo usuário;)
2. (Inserir um novo corpus de dados;)
3. (Adicionar uma lista de *stopwords*;))
4. (Digitar um novo texto para análise;)
5. (Analisar e aplicar os padrões linguísticos definidos previamente e classificar as palavras nas seguintes polaridades: Triste, Feliz, Irritado, Tenso, Animado, Entendiado, Cansado, Sonolento, Sereno e Encantando;)
6. (Analisar através de uma tabela se a ferramenta classificou corretamente as palavras.)

4.6 Rigor da Pesquisa

De acordo com Thomas e Hatchuel (2009), uma pesquisa confiável é aquela que se preocupa com a relevância e como o rigor, o qual deve estar presente em todas as fases da DSR. Em busca de apresentar o rigor na pesquisa é descrito os experimentos e métodos utilizados na presente pesquisa.

Segundo Hevner (2007), para obter respostas a questões de conhecimento, pode-se obter tanto por aplicação de artefato em um contexto, por meio de pesquisa empírica, pesquisa analítica, como também por métodos matemáticos. Na presente pesquisa, foi feita a avaliação analítica do artefato com a aplicação da medida de acurácia e precisão, utilizada com predominância nos trabalhos descritos na revisão sistemática de literatura (seção 3.2).

Para testar a precisão da ferramenta em classificar um novo conteúdo, primeiramente, define-se em quantos grupos o corpus será dividido, em seguida qual o classificador será utilizado (Naive Bayes), então, pode-se executar o teste com uma nova frase com cada um dos grupos formado e, ao final, obtém-se o percentual de precisão, computando a média aritmética com todas as precisões dos grupos testados. O resultado final se dá pela média da classificação para o corpus.

4.7 Avaliação

Esta subseção tem por objetivo descrever o processo de verificação da consistência do modelo ora proposto, foram realizados testes para avaliar a qualidade dos resultados obtidos pela ferramenta de classificação. Para tanto, foram utilizados dois conjuntos de textos (*corpus*) que possuem classificação prévia. Um *corpus* com apenas duas polaridades e um outro com seis polaridades, conforme descrito na questão de pesquisa dessa tese.

4.7.1 Experimentos

Os experimentos foram conduzidos em um conjunto de dados sobre a redução da maioria penal no Brasil (experimento 1) e em um outro sem um tema específico (experimento 2). O objetivo do (experimento 1) foi avaliar a eficácia da ferramenta criada nessa tese e das técnicas de classificação de sentimentos sobre um tema específico, que possuisse apenas duas polaridades, no caso, (favor ou contra) à redução da maioria penal. Já o (experimento 2) tem como objetivo, avaliar a precisão da ferramenta quando é necessário analisar mais de duas polaridades, no caso desse trabalho, foram realizado os testes com 6 polaridades, conforme descrito na questão de pesquisa.

4.7.1.1 Experimento 1

O *corpus* contendo apenas duas polaridades foi montado a partir do site de mídia social *Twitter*, utilizando uma *API* oferecida pela própria empresa para estes fins. Utilizamos palavras-chave para que fosse retornado somente mensagens relacionadas ao contexto da pesquisa. Um exemplo, foi a *hashtag* maioridade penal. Dessa forma, coletou-se aproximadamente 2500 *tweets*. Após a coleta e armazenamento, foi necessário fazer a filtragem e a classificação manual, rotulando cada *tweet* entre as características (favor e contra).

Para a rotulagem, seguimos o modelo de [Pang, Lee e Vaithyanathan \(2002\)](#), que não consideraram *tweets* neutros, visto que não há na literatura um entendimento sobre quais seriam as características padrão de textos classificados desta forma. Após a rotulagem, foi criada uma base com 1800 dados, divididos e rotulados em 50% a favor e 50% contra a redução da maioria penal. O processo de rotulagem foi realizado por duas pessoas, um com experiência em linguística e outro com experiência em linguística computacional. Após o processo de rotulagem individual, em função de alguns conflitos em relação a classificação (um mesmo texto recebendo rotulagem diferente), houve um processo de rotulagem “consensual” para resolver as divergências encontradas.

De posse dos resultados, (tabela 10) cada notícia foi lida individualmente e confrontada com a possível impressão causada a um leitor do texto. Dos 200 comentários avaliados nessa etapa, 156 tiveram suas polaridades identificadas corretamente pelo método. Isso equivale a uma taxa de acerto de aproximadamente 78%. Tal resultado pode ser considerado bom, visto que é equivalente

às taxas de acerto obtidas em implementações semelhantes como no trabalho de Pang, Lee e Vaithyanathan (2002) que encontraram uma acurácia de 83% para o problema de classificação de polaridade em comentários de filmes.

Tabela 10 – Acurácia do classificador para cada uma das categorias testadas

Polaridade	Acurácia	Precisão
Favor	0,83	0.80
Contra	0,74	0.78
Média	0,78	0.83

Fonte: Autor

4.7.1.2 Erros em classificações

Mesmo alcançando um percentual de 78%, o que já está próximo ao alcançado em outros trabalhos e, muito próximo a capacidade humana de avaliar um texto subjetivamente, alguns erros foram detectados nas classificações.

Exemplo 1: “Eu acho que um adolescente não pode ser responsabilizado pelos seus atos”, “favor”.

O sentimento real é contra a Redução da Maioridade Penal, mas o algoritmo *Naive Bayes* previu para a classe à (favor). O modelo tem uma probabilidade sobre a palavra "adolescente" para a classe (contra), que não leva em conta a palavra negativa "não" depois dela. Isso nos leva a reflexão de que são necessárias muitas frases de cada sentimento para obter um resultado mais preciso.

4.7.1.3 Experimento 2

O *corpus* contendo seis polaridades também foi montado a partir do site de mídia social *Twitter*. Coletou-se aproximadamente 1600 *tweets*, armazenados em formato texto, respeitando as polaridades definidas na questão de pesquisa desse trabalho, a qual é dividida em seis emoções. Dos 1600 *tweets*, foram separados 1200 para a fase de treinamento e 400 para a fase de teste. Esse pequeno conjunto de dados, composto por apenas 1600 mensagens curtas, deve-se ao fato que estudos anteriores, como o de Banea et al. (2008) mostraram que pequenos corpora são capazes de atingir um alto índice de acertos em classificação.

Após a coleta e armazenamento, foi necessário fazer a filtragem e a classificação manual, rotulando cada um dos 1200 *tweets* entre as emoções definidas na questão de pesquisa. Após este processo, foi criada uma base de dados com 1200 frases, divididas e rotuladas proporcionalmente igual para cada emoção.

Analisando os dados obtidos na tabela 11, verifica-se que o classificador obteve um resultado inferior a 50% ao classificar textos referentes as emoções "Entendiado", "Irritado" e "Serenio".

Tabela 11 – Acurácia do classificador para cada uma das categorias testadas

Polaridade	Acurácia	Precisão
Angustiado	0,59	0.61
Animado	0,67	0.65
Cansado	0,63	0.66
Entendiado	0,47	0.45
Encantado	0,55	0.57
Feliz	0,59	0.61
Irritado	0,42	0.39
Sonolento	0,53	0.52
Sereno	0,31	0.27
Satisfeito	0,64	0.60
Triste	0,71	0.73
Tenso	0,50	0.49
Média	0,55	0.54

Fonte: Autor

Acredita-se que a razão para tal resultado está bem abaixo do mínimo esperado de uma classificação humana, variando entre 72% [Wiebe e Riloff \(2005\)](#) a 85% [Golden \(2011\)](#), deve-se ao fato de que os textos relacionados a essas categorias possuem construções mais difíceis de serem analisadas automaticamente. Além disso, algumas dessas mensagens estão fortemente ligadas ao contexto a que estão relacionadas, dificultando ainda mais a classificação, tanto manual como automática. Outro fator importante para essa análise é a quantidade de dados para o treinamento, em que, cada emoção contia apenas 200 mensagens treinadas na base de dados.

4.7.2 Análise dos experimentos

Observa-se para os conjuntos de dados analisados, que tanto para situações de taxa de acerto alta (experimento 1) e baixa (experimento 2), a acurácia e a precisão são baixas com uma base de dados com poucos dados para treinamento. Isso vai contra as achados de [Banea et al. \(2008\)](#) que mostraram em seu trabalho bons resultados com um corpora pequeno, na língua inglesa.

Os resultados encontrado nessa dissertação condizem mais com a idéia de que uma grande quantidade de termos similares ajuda no processo de classificação, embora, isso torna o processo mais lento, causando até *overstemming* [Abbasi, Chen e Salem \(2008\)](#). A interpretação mais plausível desses resultados é que uma base de dados com poucos termos enfraquecem a força dos atributos que podem auxiliar a tarefa de classificação, comprometendo a exatidão. Assim, um fator importante que não foi enfatizado nessa tese, é a quantidade mínima e máxima de termos para o processo de treinamento que, poderia, se trabalho adequadamente, fornecer ganhos ao processos subsequentes.

4.8 Síntese das contribuições gerais desta pesquisa

Hevner (2007) elaborou as diretrizes da relevância das pesquisas com DSR (Seção 4), as quais foram descritas e relacionadas com este trabalho. A quarta diretriz descreve a necessidade de complementar conhecimentos à base já existente ou aplica-lós de novas maneiras. Finalmente, uma pesquisa conduzida pela DSR deve prover contribuições nas áreas específicas dos artefatos desenvolvidos.

Aproveitando-se desta diretriz, elaborou-se uma síntese das contribuições gerais da presente pesquisa:

1. **adoção da metodologia *design science research*** (Seção 4);
2. **elaboração de uma questão de pesquisa que oriente a parte prática da pesquisa:** recorreu-se à confluência de outras abordagens: revisão sistemática de literatura (Seção 3.2);
3. **concepção de um modelo como contribuição ao conhecimento:** valeu-se de um artefato para a solução de problema real, o modelo de mineração de textos, suas variáveis e relações conectadas (Seção 4.5.2);
4. **geração de conhecimento através das diretrizes de Hevner (2007):** sistematizou-se nas sete etapas do ciclo, os problemas práticos .
5. **investigação, avaliação e validações do problema:** realizou-se investigações, avaliações e validações do problema de pesquisa, quando emergiram-se, (Seções 3.2, 4.7.1, 4.7.2) ;

Finda-se aqui a discussão deste trabalho. O próximo Capítulo contém as principais conclusões, as limitações do estudo e as sugestões para trabalhos futuros.

4.9 Conclusão

Na presente pesquisa, quanto à Ciência do *Design*, foi elaborado um artefato, que é um modelo matemático de sistema para detecção automática de sentimentos em site de mídias sociais. Após ampla revisão bibliográfica, analisando, principalmente, as técnicas para análise de sentimentos em textos, foi apresentado um modelo que atende de forma satisfatória os requisitos propostos.

Embora o modelo tenha sido desenvolvido com o objetivo de classificar dados no idioma Português, a sua utilização em outros idiomas é passível sem grande esforço, pois, toda parte de pré-processamento funciona através de arquivos externos, do formato .csv, como no caso da lista de *stopwords*.

Além disso, a solução apresentada utiliza a biblioteca gratuita NLTK, vista na seção 4.5.1.4, que permite estudo e melhor entendimento do processo a custo zero. Diante dessa perspectiva, do

mesmo modo como foi utilizado o *stemmer* para língua portuguesa, poderia utilizar-se de um *stemmer* para outro idioma, sem nenhuma complicação.

O trabalho estruturou-se em três objetivos específicos. O primeiro compreendia identificar os métodos e técnicas utilizadas na análise de sentimentos em textos. Este objetivo foi alcançado, conforme apresentado na seção 3.3, por meio dos resultados da revisão sistemática de literatura. O segundo e o terceiro objetivo fez-se a escolha dos métodos e bibliotecas para auxiliar no desenvolvimento da ferramenta. Estes objetivos foram plenamente alcançados, sendo apresentado um modelo que compreende todas as atividades do processo para analisar sentimentos sobre textos extraídos de mídias sociais, foi apresentado todos os requisitos necessários à implementação desse sistema.

Concluída a implementação do método, iniciou-se os testes sobre dois experimentos para validar o modelo elaborado. Para isso, foi necessário coletar e montar duas bases de dados, uma contendo apenas duas polaridades e uma outra com seis polaridades. Optou-se por coletar mensagens do Twitter, por ser um site de mídia social com vasto conteúdo de linguagem natural. A ideia de testar o modelo com uma base de dados com mais de duas polaridades, surgiu da escassez de ferramentas em português do Brasil explorando métodos e técnicas da análise de sentimentos em textos.

Através dos experimentos realizados, mostrou-se que é possível classificar sentimentos em textos, escrito em português, de forma automatizada, através de algoritmos de mineração de textos. Embora a eficácia comprovada da ferramenta, no (experimento 2) ainda seja muito inferior a capacidade humana de avaliar um texto subjetivamente, variando entre 74% à 83%.

Percebeu-se através dos resultados alcançados tanto na acurácia como na precisão que a melhoria no processo de classificação está intimamente ligado à quantidade de dados na fase de treinamento.

Por fim, a solução atendeu plenamente aos objetivos propostos. Os testes e avaliações apresentados mostraram a aplicabilidade da solução proposta para o problema de classificação de sentimentos em textos para o Idioma Português do Brasil.

4.10 Limitações do trabalho

Nesse trabalho assumiu-se como premissa que seria possível classificar textos oriundos de sites de mídias sociais. O desafio era verificar, de um modo estatisticamente válido, quais e que tipos de características seria mais profícuas para implementar na ferramenta de classificação automática. Assumi-se também, que a ferramenta AEMiner poderia suportar boa parte dessas características, com a vantagem de executar em um ambiente web, trabalhando com independência de corpus e que possuísse um ambiente de fácil aprendizagem para não especialistas em estruturação de dados.

Genuinamente, desde longa data, é possível extrair, identificar e classificar textos automaticamente, independente do objetivo que se tenha, e os buscadores *stopwords*(*Google, Yahoo*) nos mostra isso diariamente. Considerando-se uma busca refinada nessas ferramentas, para além de um documento contendo apenas uma palavra ou expressão, dentre inúmeras possibilidades de separações já passíveis automaticamente, é possível separar textos extensos dos textos pequenos, textos com títulos dos sem títulos.

Essas separações se tornam inteligível caso haja uma etiqueta consideravelmente distintiva nos textos que se acessa ou se estiver visível a, hoje modesta, informação do recurso Wordcount (contador de palavras) ao lado do arquivo/texto. Por essa ótica, é possível comparar grandes corpora entre si, apenas analisando as suas características lexicais e os tipos de gêneros de textos neles contidos, desde que os dados estejam listados.

Sabendo-se disso, na presente pesquisa a questão de pesquisa principal foi: Quais os requisitos e os componentes de um sistema informatizado que seja capaz de analisar sentimentos em textos extraído de mídias sociais, utilizando as polaridades: Angustiado, Animado, Cansado, Entendiado, Encantado, Feliz, Irritado, Sonolento, Sereno, Satisfeito, Triste e Tenso?

A junção entre os princípios da PLN e as técnicas para análise de sentimentos mostra que é possível selecionar determinadas características em um texto e manipulá-las de um modo eficiente, para fins de classificação automática. Nessa pesquisa, navegou-se entre o mercado de software e a literatura, buscando o que já existe de ferramentas para a tarefa de classificação.

As características desenvolvidas na ferramenta AEMiner - majoritariamente coesivas com a literatura - forneceram informações a serem comparados entre os textos. Essas informações, por sua vez, foram processadas no artefato desenvolvido, com a finalidade de estabelecer uma correlação estatística e automática na classificação. O resultados mostraram que a solução atendeu plenamente aos objetivos propostos.

Evidentemente, enfrentou-se limitações na proposição de um modelo de classificação automática de textos. A primeira limitação foi a ausência de um *corpus* maior, o que, como descritos na análise dos experimentos (seção 4.7.2) , inibe o poder do algoritmo em classificar um texto corretamente. A ferramenta AEMiner exige um cuidadoso trabalho prévio com os textos a ela submetidos, especialmente a conferência de pontuações.

A segunda limitação está ligada a natureza do modelo informatizado, o AEMiner, que produz uma série de informações sobre o *corpus* classificado, mas, ainda sim, são considerados elementos superficiais, tais como termos mais frequentes, contagens de palavras, tamanho de sentenças. Além disso, a ferramenta dispõem apenas de duas métricas, ficando um pouco comprometida, caso seja necessário analisar os resultados obtidos sobre outras métricas, tais como *F-Mensuare*, *stopwords* e *Recall*.

A terceira limitação, e talvez a maior delas, deriva-se justamente da tentativa de criar procedimentos padronizados através de uma tarefa complexa e subjetiva: a avaliação humana. Diante de

textos extraídos do site de mídia social *Twitter*, em que os recursos linguísticos presentes nos textos possuem construções mais difíceis de serem analisadas, apontar itens específicos capazes de diferenciar os textos foi uma tarefa onerosa. Além disso, algumas dessas mensagens estão fortemente ligadas ao contexto a que estão relacionadas, dificultando ainda mais a classificação. Contudo, apesar das limitações, acredita-se que a ideia metodológica proposta nessa pesquisa tem grande potencial de aproveitamento. No entanto, o caráter experimental dessa dissertação, faz-se necessário averiguar o alcance do artefato criado para um universo maior, tanto para outros domínios de negócio, como para outras línguas, algo a ser feito em larga escala e em trabalhos futuros.

4.11 Trabalhos futuros

Como trabalhos futuros, vislumbra-se a implementação de uma avatar no modelo, para que possa ser verificado através de uma animação as expressões faciais dos sentimentos contidos nas mensagens. Pode-se ainda avançar numa linha de pesquisa sobre a aplicação funcionando em ambientes text-to-speech, de forma a conciliar a expressão facial com a entonação de uma voz. Pretende-se também, ampliar os textos do corpus para que seja alcançados melhores resultados na precisão do algoritmo. Outra extensão prevista é a independência do sistema quanto a base de treinamento, isso pode ser obtido com pequenas modificações no sistema, já que os módulos são independentes, com isso seria possível aplicar o processo de classificação em diferentes domínios de negócios de forma simultânea. Através dos experimentos foi possível identificar a influência da quantidade de termos na acurácia da classificação de documentos. Um estudo mais aprofundado poderia determinar a relação da quantidade de termos, de forma a maximar a eficiência do sistema.

Outro ponto que pode ser investigado refere-se à utilização de diferentes medidas estatísticas na classificação da polaridade de tweets, para obter as frequências de coocorrência entre palavras. Dentre essas medidas, pode-se citar Latent Semantic Analysis (LSA) [Landauer e Dumais \(1997\)](#) e Normalized Web Distance (NWD) ([CILIBRASI; VITANYI, 2009](#)).

Embora esteja muito além do escopo deste estudo, pode ser interessante usar regressão em vez de classificação. A regressão prediz um valor entre, por exemplo, -1 e +1, e fornece mais espaço para a predição.

Pesquisas na análise de sentimentos já produziram algumas aplicações inovadoras, mas ainda há um longo caminho a percorrer antes que os computadores compreendam verdadeiramente essas duas coisas que todos os seres humanos têm em comum: emoções e linguagem.

Cabe, ainda, ressaltar que todos os dados analisados através dos corpus foram corrigidos por humanos e classificados por eles em níveis pré-determinados e, essa avaliação não foi questionada. Sendo, através dela que os testes, medidas e escores estatísticos foram executados, dispondo

de uma confiabilidade máxima do quanto os dados da pesquisa se aproximavam das medidas humanas. Acontece que a correção humana é bastante subjetiva, repletos de idiossincrasias. Então, o que buscou-se, foi atingir níveis de classificação humana descritos na literatura, conforme enfatizado em vários momentos dessa dissertação. Talvez fosse interessante desprezar os avaliadores humanos, assim, outros fenômenos poderiam ser observados e que não foram pela necessidade de trabalhar com os níveis pré-determinados.

Enfim, há muito o que fazer. Viu-se que, a aplicação dos algoritmos de PLN para o português do Brasil engatinha na produção de ferramentas, e para impulsionar a área, nós, especialistas em sistemas de informação, somos fundamentais.

Referências

- ABBASI, A.; CHEN, H.; SALEM, A. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, ACM, v. 26, n. 3, p. 12, 2008.
- AHMAD, T.; DOJA, M. N. Ranking system for opinion mining of features from review documents. *IJCSI International Journal of Computer Science Issues*, Citeseer, v. 9, n. 4, p. 1694–0814, 2012.
- ANANIADOU, S.; MCNAUGHT, J. *Text mining for biology and biomedicine*. [S.l.]: Citeseer, 2006.
- ARANHA, C. N.; VELLASCO, M. Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional. *Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ*, 2007.
- BACH, N. X.; PHUONG, T. M. Leveraging User Ratings for Resource-poor Sentiment Classification. *Procedia Computer Science*, v. 60, p. 322–331, 2015. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050915022619>>.
- BAEZA-YATES, R.-N. *Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval. Chapter 3*. [S.l.]: Addison Wesley, 1999.
- BAFNA, K.; TOSHNIWAL, D. Feature based Summarization of Customers' Reviews of Online Products. *Procedia Computer Science*, v. 22, p. 142–151, 2013. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050913008831>>.
- BAIR, E.; TIBSHIRANI, R. Machine learning methods applied to dna microarray data can improve the diagnosis of cancer. *ACM SIGKDD Explorations Newsletter*, ACM, v. 5, n. 2, p. 48–55, 2003.
- BANEA, C. et al. Multilingual subjectivity analysis using machine translation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. [S.l.], 2008. p. 127–135.
- BARAWI, M. H.; SENG, Y. Y. Evaluation of Resource Creations Accuracy by Using Sentiment Analysis. *Procedia - Social and Behavioral Sciences*, v. 97, p. 522–527, nov. 2013. ISSN 1877-0428. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877042813037130>>.
- BARBOSA, R. R. L.; SÁNCHEZ-ALONSO, S.; SICILIA-URBAN, M. A. Evaluating hotels rating prediction based on sentiment analysis services. *Aslib Journal of Information Management*, v. 67, n. 4, p. 392–407, jul. 2015. ISSN 2050-3806. Disponível em: <<http://www.emeraldinsight.com/doi/10.1108/AJIM-01-2015-0004>>.
- BARNES, N. G.; LESCAULT, A. M. Social media adoption soars as higher-ed experiments and reevaluates its use of new communications tools. *Center for Marketing Research. University of Massachusetts Dartmouth, North Dartmouth, MA*, 2011.
- BEAZLEY, D. M. *Python essential reference*. [S.l.]: Sams Publishing, 2006.

- BERRY, M. W.; KOGAN, J. *Text mining: applications and theory*. [S.l.]: John Wiley & Sons, 2010.
- BHASKAR, J.; SRUTHI, K.; NEDUNGADI, P. Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining. *Procedia Computer Science*, v. 46, p. 635–643, 2015. ISSN 1877-0509. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050915001763>.
- BILAL, M. et al. Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. *Journal of King Saud University - Computer and Information Sciences*, jul. 2015. ISSN 1319-1578. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1319157815001330>.
- BURIOL, T. M.; ARGENTA, M. A. Acelerando o desenvolvimento e o processamento de análises numéricas computacionais utilizando python e cuda. *Métodos Numéricos e Computacionais em Engenharia-CMNE CILAMCE*, 2009.
- CHAKRABARTI, S. *Mining the Web: Discovering knowledge from hypertext data*. [S.l.]: Elsevier, 2002.
- CHEN, H. Knowledge management systems: a text mining perspective. Knowledge Computing Corporation, 2001.
- CHOWDHURY, G. G. Natural language processing. *Annual review of information science and technology*, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003.
- CILIBRASI, R. L.; VITANYI, P. Normalized web distance and word similarity. *arXiv preprint arXiv:0905.4039*, 2009.
- CUNHA, C.; CINTRA, L. F. L. et al. *Nova gramática do português contemporâneo*. [S.l.]: Nova Fronteira Rio de Janeiro, 1985.
- DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: ACM. *Proceedings of the 12th international conference on World Wide Web*. [S.l.], 2003. p. 519–528.
- DRESCH, A.; LACERDA, D. P.; JÚNIOR, J. A. V. A. *Design science research: método de pesquisa para avanço da ciência e tecnologia*. [S.l.]: Bookman Editora, 2015.
- EFRON, M. Cultural orientation: Classifying subjective documents by cociation analysis. In: *In AAAI Fall Symposium on Style and Meaning in*. [S.l.: s.n.], 2004.
- FELDMAN, R.; SANGER, J. *The text mining handbook: advanced approaches in analyzing unstructured data*. [S.l.]: Cambridge University Press, 2007.
- FROELICH, J.; ANANYAN, S. Decision support via text mining. In: *Handbook on Decision Support Systems 1*. [S.l.]: Springer, 2008. p. 609–635.
- GIATSOGLU, M. et al. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, Elsevier, v. 69, p. 214–224, 2017.
- GODBOLE, N.; SRINIVASAIAH, M.; SKIENA, S. Large-scale sentiment analysis for news and blogs. *ICWSM*, v. 7, n. 21, p. 219–222, 2007.

- GOEURLOT, L. et al. Sentiment lexicons for health-related opinion mining. In: ACM. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. [S.l.], 2012. p. 219–226.
- GOLDEN, P. Write here, write now. *Research*, May, 2011.
- GOMES, H.; NETO, M. de C.; HENRIQUES, R. Text mining: Sentiment analysis on news classification. In: IEEE. *2013 8th Iberian Conference on Information Systems and Technologies (CISTI)*. [S.l.], 2013. p. 1–6.
- GONÇALVES, P. et al. Comparing and combining sentiment analysis methods. In: ACM. *Proceedings of the first ACM conference on Online social networks*. [S.l.], 2013. p. 27–38.
- GONÇALVES, T. et al. Analysing part-of-speech for portuguese text classification. In: *Computational Linguistics and Intelligent Text Processing*. [S.l.]: Springer, 2006. p. 551–562.
- HADANO, M.; SHIMADA, K.; ENDO, T. Aspect Identification of Sentiment Sentences Using A Clustering Algorithm. *Procedia - Social and Behavioral Sciences*, v. 27, p. 22–31, 2011. ISSN 1877-0428. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877042811024062>>.
- HAMMAMI, M.; GUERMAZI, R.; HAMADOU, A. B. Automatic violent content web filtering approach based on the KDD process. *International Journal of Web Information Systems*, v. 4, n. 4, p. 441–464, nov. 2008. ISSN 1744-0084. Disponível em: <<http://www.emeraldinsight.com/doi/abs/10.1108/17440080810919486>>.
- HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques: concepts and techniques*. [S.l.]: Elsevier, 2011.
- HEVNER, A. R. A three cycle view of design science research. *Scandinavian journal of information systems*, v. 19, n. 2, p. 4, 2007.
- HU, M.; LIU, B. Mining and summarizing customer reviews. In: ACM. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2004. p. 168–177.
- INDURKHYA, N.; DAMERAU, F. J. *Handbook of natural language processing*. [S.l.]: CRC Press, 2010.
- JACKSON, P.; MOULINIER, I. *Natural language processing for online applications: Text retrieval, extraction and categorization*. [S.l.]: John Benjamins Publishing, 2007.
- JAIN, V. K.; KUMAR, S. An Effective Approach to Track Levels of Influenza-A (H1n1) Pandemic in India Using Twitter. *Procedia Computer Science*, v. 70, p. 801–807, 2015. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050915032846>>.
- JUNIOR, J. R. C. *Desenvolvimento de uma Metodologia para Mineração de Textos*. Tese (Doutorado) — PUC-Rio, 2007.
- KANSAL, H.; TOSHNIWAL, D. Aspect based Summarization of Context Dependent Opinion Words. *Procedia Computer Science*, v. 35, p. 166–175, 2014. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050914010618>>.

KECHAOU, Z. et al. A novel system for video news' sentiment analysis. *Journal of Systems and Information Technology*, v. 15, n. 1, p. 24–44, mar. 2013. ISSN 1328-7265. Disponível em: <<http://www.emeraldinsight.com/doi/abs/10.1108/13287261311322576>>.

KIM, Y.; STREET, W. N.; MENCZER, F. Feature selection in unsupervised learning via evolutionary search. In: ACM. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2000. p. 365–369.

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, v. 33, n. 2004, p. 1–26, 2004.

KLEMANN, M.; REATEGUI, E.; LORENZATTI, A. O emprego da ferramenta de mineração de textos sobek como apoio à produção textual. In: *Anais do Simpósio Brasileiro de Informática na Educação*. [S.l.: s.n.], 2009. v. 1, n. 1.

KLEMANN, M.; REATEGUI, E.; RAPKIEWICZ, C. Análise de ferramentas de mineração de textos para apoio a produção textual. In: *Anais do Simpósio Brasileiro de Informática na Educação*. [S.l.: s.n.], 2011. v. 1, n. 1.

K.M., A. K. et al. A Multimodal Approach To Detect User's Emotion. *Procedia Computer Science*, v. 70, p. 296–303, 2015. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050915032603>>.

KOTHARI, A. A.; PATEL, W. D. A Novel Approach Towards Context Based Recommendations Using Support Vector Machine Methodology. *Procedia Computer Science*, v. 57, p. 1171–1178, 2015. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050915019377>>.

KUMAR, K. M. A. et al. Analysis of users' Sentiments from Kannada Web Documents. *Procedia Computer Science*, v. 54, p. 247–256, 2015. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050915013538>>.

LANDAUER, T. K.; DUMAIS, S. T. The latent semantic analysis theory of acquisition. *Induction and Representation of Knowledge* 1997, 1997.

LANDEGHEM, S. V. et al. High-Precision Bio-Molecular Event Extraction from Text Using Parallel Binary Classifiers. *Computational Intelligence*, v. 27, n. 4, p. 645–664, nov. 2011. ISSN 08247935. Disponível em: <<http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=67460992&site=ehost-live>>.

LI, S.; LEE, S. Y. M.; HUANG, C.-R. Corpus construction on polarity shifting in sentiment analysis. In: SPRINGER. *Workshop on Chinese Lexical Semantics*. [S.l.], 2013. p. 625–634.

LIDDY, E. *Encyclopedia of library and information science*. new york: Marcel dekker. Inc.[Links], 2001.

LIDDY, E. D.; MCCRACKEN, N. J. Hands-on nlp for an interdisciplinary audience. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. [S.l.], 2005. p. 62–68.

LIU, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

LONGHI, M. et al. Um framework para tratamento do léxico afetivo a partir de textos disponibilizados em um ambiente virtual de aprendizagem. *RENOTE*, v. 8, n. 2, 2010.

LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*. [S.l.], 2002. p. 63–70.

LOPES, I. L. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. *Ciência da informação*, SciELO Brasil, v. 31, n. 1, p. 41–52, 2002.

MANNING, C. D. et al. *Introduction to information retrieval*. [S.l.]: Cambridge university press Cambridge, 2008.

MARRESE-TAYLOR, E. et al. Identifying Customer Preferences about Tourism Products Using an Aspect-based Opinion Mining Approach. *Procedia Computer Science*, v. 22, p. 182–191, 2013. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050913008879>>.

MARTINS, J. Classificação de páginas na internet. *Trabalho de Conclusão (Mestrado)*. Instituto de Ciências Matemáticas e de Computação. USP. São Carlos, 2003.

MITCHELL, T. M. *Machine learning*. WCB. [S.l.]: McGraw-Hill Boston, MA:, 1997.

MOOERS, C. Zatacoding applied to mechanical organization of knowledge. *American Documentation*, v. 2, n. 1, 1951.

MORENO-ORTIZ, A.; FERNÁNDEZ-CRUZ, J. Identifying Polarity in Financial Texts for Sentiment Analysis: A Corpus-based Approach. *Procedia - Social and Behavioral Sciences*, v. 198, p. 330–338, jul. 2015. ISSN 1877-0428. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877042815044523>>.

NASUKAWA, T.; NAGANO, T. Text analysis and knowledge mining system. *IBM systems journal*, IBM, v. 40, n. 4, p. 967–984, 2001.

ORENGO, V. M.; HUYCK, C. R. A stemming algorithm for the portuguese language. In: *spire*. [S.l.: s.n.], 2001. v. 8, p. 186–193.

O'RIORDAN, C.; SORENSEN, H. Information filtering and retrieval: An overview. *citeseer.nj.nec.com/483228.html*, 1997.

PAK, A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In: *LREC*. [S.l.: s.n.], 2010. v. 10, p. 1320–1326.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, Now Publishers Inc., v. 2, n. 1-2, p. 1–135, 2008.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. [S.l.], 2002. p. 79–86.

- PEFFERS, K. et al. The design science research process: a model for producing and presenting information systems research. In: *Proceedings of the first international conference on design science research in information systems and technology (DESRIST 2006)*. [S.l.: s.n.], 2006. p. 83–106.
- PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. *Linguagens Especializadas em Corpora: modos de dizer e interfaces de pesquisa*. [S.l.]: EDIPUCRS, 2010.
- PRABOWO, R.; THELWALL, M. Sentiment analysis: A combined approach. *Journal of Informetrics*, Elsevier, v. 3, n. 2, p. 143–157, 2009.
- PREETHI, P. G.; UMA, V.; KUMAR, A. Temporal Sentiment Analysis and Causal Rules Extraction from Tweets for Event Prediction. *Procedia Computer Science*, v. 48, p. 84–89, 2015. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050915006638>>.
- RODRIGUES, R. G. et al. Sentihealth-cancer: uma ferramenta de análise de sentimento para ajudar a detectar o humor de pacientes de câncer em uma rede social online. Universidade Federal de Goiás, 2016.
- ROMAN, N. *Emoção ea Sumarização Automática de Diálogos*. Tese (Doutorado) — PhD thesis, Instituto de Computação–Universidade Estadual de Campinas, Campinas, Sao Paulo, 2007.
- SÆTRE, R. et al. Semantic annotation of biomedical literature using google. In: *Computational Science and Its Applications–ICCSA 2005*. [S.l.]: Springer, 2005. p. 327–337.
- SALTON, G. Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*, 1989.
- SANCHEZ-MONZON, J.; PUTZKE, J.; FISCHBACH, K. Automatic Generation of Product Association Networks Using Latent Dirichlet Allocation. *Procedia - Social and Behavioral Sciences*, v. 26, p. 63–75, 2011. ISSN 1877-0428. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877042811023901>>.
- SCHMID, H. Part-of-speech tagging with neural networks. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 15th conference on Computational linguistics-Volume I*. [S.l.], 1994. p. 172–176.
- SEBASTIANI, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, ACM, v. 34, n. 1, p. 1–47, 2002.
- SHAHANA, P. H.; OMMAN, B. Evaluation of Features on Sentimental Analysis. *Procedia Computer Science*, v. 46, p. 1585–1592, 2015. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050915001520>>.
- SILVA, N. G. R. da. *BestChoice: Classificação de Sentimento em Ferramentas de Expressão de Opinião*. Tese (Doutorado) — Tese de graduação, Universidade Federal de Pernambuco, Recife, 2010. 7, 17, 2010.
- STOYANOV, V.; CARDIE, C.; WIEBE, J. Multi-perspective question answering using the opqa corpus. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. [S.l.], 2005. p. 923–930.

- STROUD, D. Social networking: An age-neutral commodity—social networking becomes a mature web application. *Journal of Direct, Data and Digital Marketing Practice*, Nature Publishing Group, v. 9, n. 3, p. 278–292, 2008.
- TAMAMES, J.; LORENZO, V. de. EnvMine: A text-mining system for the automatic extraction of contextual information. *BMC Bioinformatics*, v. 11, p. 294–303, jan. 2010. ISSN 14712105. Disponível em: <<http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=52799161&site=ehost-live>>.
- TAN, A.-H. et al. Text mining: The state of the art and the challenges. In: *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*. [S.l.: s.n.], 1999. v. 8, p. 65–70.
- TURBAN, E. et al. *Tecnologia da Informação para Gestão: Transformando os Negócios na Economia Digital 6ed.* [S.l.]: Bookman, 2010.
- TURNER, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 40th annual meeting on association for computational linguistics*. [S.l.], 2002. p. 417–424.
- VALSAMIDIS, S. et al. A Framework for Opinion Mining in Blogs for Agriculture. *Procedia Technology*, v. 8, p. 264–274, 2013. ISSN 2212-0173. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S2212017313000984>>.
- VIDYA, N. A.; FANANY, M. I.; BUDI, I. Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers. *Procedia Computer Science*, v. 72, p. 519–526, 2015. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050915036200>>.
- VIECHNICKI, P. A performance evaluation of automatic survey classifiers. In: SPRINGER. *International Colloquium on Grammatical Inference*. [S.l.], 1998. p. 244–256.
- WEICHSELBRAUN, A.; GINDL, S.; SCHARL, A. Enriching semantic knowledge bases for opinion mining in big data applications. *Knowledge-Based Systems*, v. 69, p. 78–85, out. 2014. ISSN 0950-7051. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0950705114001695>>.
- WEISS, S. M. et al. *Text mining: predictive methods for analyzing unstructured information*. [S.l.]: Springer Science & Business Media, 2010.
- WIEBE, J.; RILOFF, E. Creating subjective and objective sentence classifiers from unannotated texts. In: *Computational Linguistics and Intelligent Text Processing*. [S.l.]: Springer, 2005. p. 486–497.
- WIEBE, J. et al. Learning subjective language. *Computational linguistics*, MIT Press, v. 30, n. 3, p. 277–308, 2004.
- WIERINGA, R. Design science as nested problem solving. In: ACM. *Proceedings of the 4th international conference on design science research in information systems and technology*. [S.l.], 2009. p. 8.
- WILSON, T. et al. Opinionfinder: A system for subjectivity analysis. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of hlt/emnlp on interactive demonstrations*. [S.l.], 2005. p. 34–35.

WILSON, T.; WIEBE, J.; HWA, R. Recognizing strong and weak opinion clauses. *Computational Intelligence*, v. 22, n. 2, p. 73–99, 2006.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2011.

WIVES, L. K.; LOH, S. Recuperação de informações usando a expansão semântica e a lógica difusa. In: *Congresso Internacional En Ingenieria Informatica, ICIE*. [S.l.: s.n.], 1998.

WOHL, A. D. Intelligent text mining creates business intelligence. *IBM Business Intelligence Solutions CD. EUA*, 1998.

YAAKUB, M. R.; LI, Y.; ZHANG, J. Integration of Sentiment Analysis into Customer Relational Model: The Importance of Feature Ontology and Synonym. *Procedia Technology*, v. 11, p. 495–501, 2013. ISSN 2212-0173. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2212017313003745>.

YU, L.; LIU, H. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research, JMLR. org*, v. 5, p. 1205–1224, 2004.

ZAGHLOUL, W.; LEE, S. M.; TRIMI, S. Text classification: neural networks vs support vector machines. *Industrial Management & Data Systems*, v. 109, n. 5, p. 708–717, maio 2009. ISSN 0263-5577. Disponível em: <http://www.emeraldinsight.com/doi/abs/10.1108/02635570910957669>.

ZHAO, M. et al. METSP: A Maximum-Entropy Classifier Based Text Mining Tool for Transporter-Substrate Identification with Semistructured Text. *BioMed Research International*, v. 2015, p. 1–7, 2015. ISSN 23146133. Disponível em: <http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=110311222&site=ehost-live>.

Apêndice A – Um apêndice

O Sistema

Os módulos da ferramenta foram desenvolvidos e implementados para funcionarem de forma integrada. A interface tem como objetivo facilitar a utilização da aplicação por usuários que não possuem conhecimentos específicos em programação.

As camadas de interface oferece ao usuário um conjunto de diálogo gráfico para definição de parâmetros e opções de execução, menus para edição e seleção dos arquivos. A figura 4.11 ilustra a tela principal da ferramenta com as opções disponíveis ao usuário.

Na tela principal, o usuário poderá criar uma nova conta, realizar uma demonstração da ferramenta sem que faça seu cadastro, fazer uma leitura da documentação de utilização da ferramenta e, ainda, entrar em contato com os desenvolvedores.

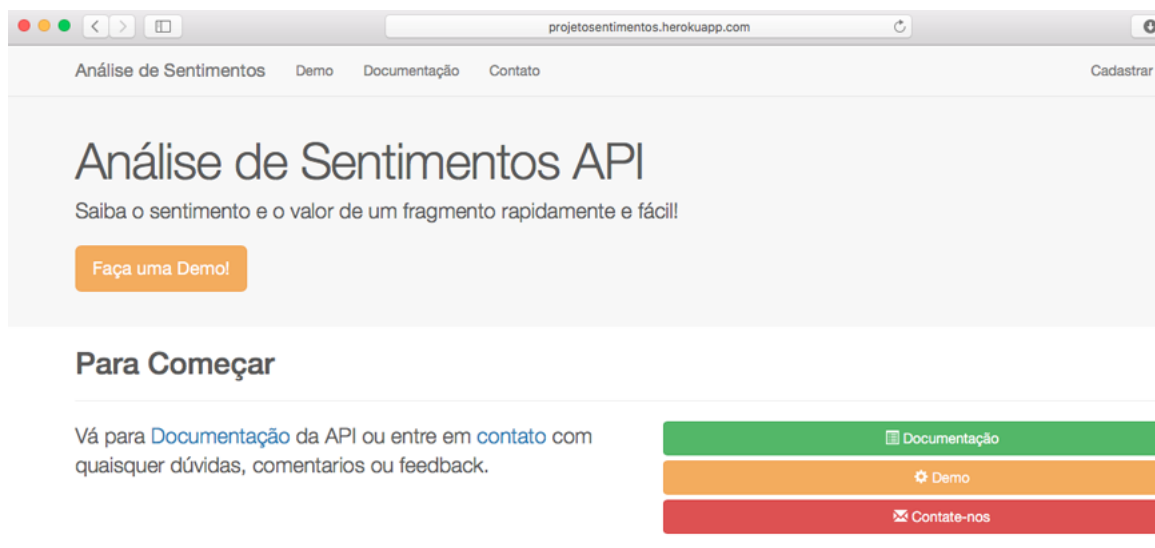


Figura 13 – Tela principal da ferramenta

Fonte: Autor

A interface mostrada na figura 14 permite que sejam definidas algumas configurações como a lista de stopwords e uma outra base de dados que não seja a própria da ferramenta. Com isso, o usuário tem liberdade de trabalhar com outros domínios de negócio.

A interface mostrada na figura 15, o usuário poderá digitar uma nova frase para testar o algoritmo de Bayes. Antes que o resultado seja mostrada na tela, são realizadas diversas atividades concomitante, dentre elas a retiradas de palavras, termos e símbolos indesejados, como por exemplo, pontuação e números, além do processo de classificação através do algoritmo Naive Bayes.

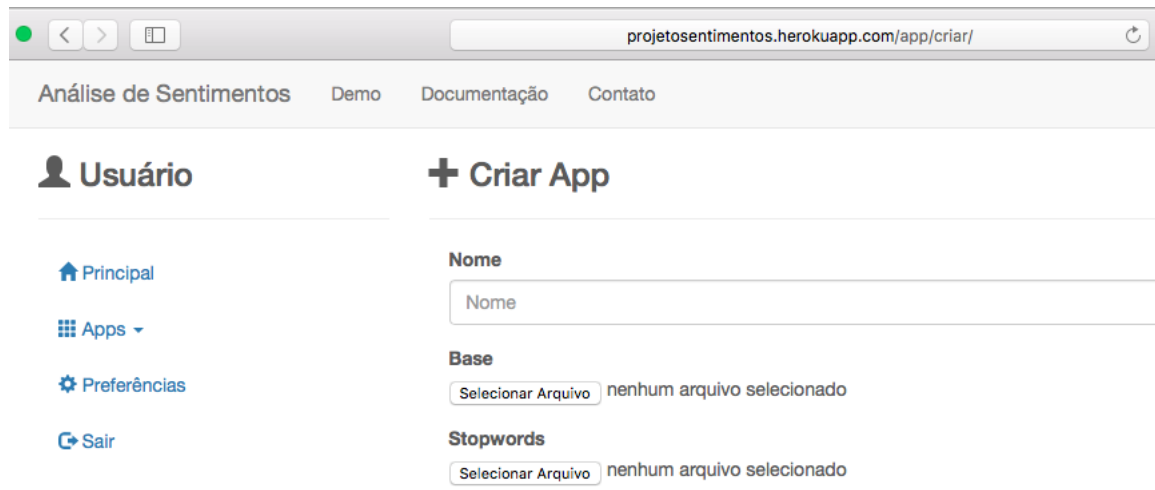


Figura 14 – Tela de inclusão da base de dados e Stopwords

Fonte: Autor

Na tela mostrada na figura 16, o usuário poderá visualizar a acurácia do algoritmo no que diz respeito às classificações. Pode-se visualizar também as últimas frases classificadas, as polaridades e o percentual de acerto. A ideia de uma visualização mais detalhada foi motivada pela necessidade de mais estudos relacionado à mineração de textos com diversas polaridades. A última tela da ferramenta, mostrada na figura 17, é disponibilizado para o usuário uma tabela contendo todas as classificações realizadas. Sendo apresentado tanto as classificações corretas como as incorretas.

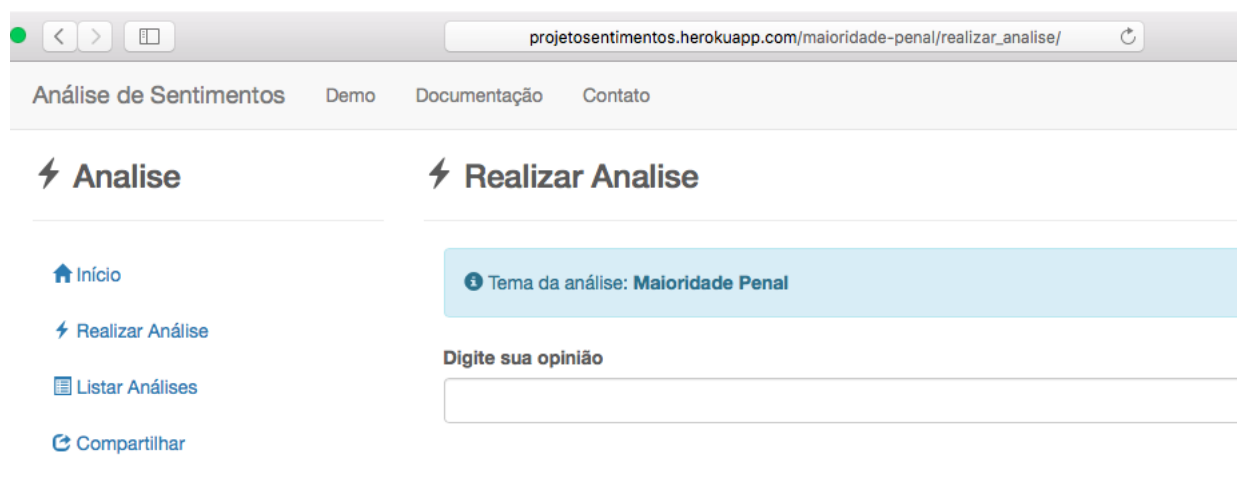


Figura 15 – Tela para realizar classificação de novas frases

Fonte: Autor

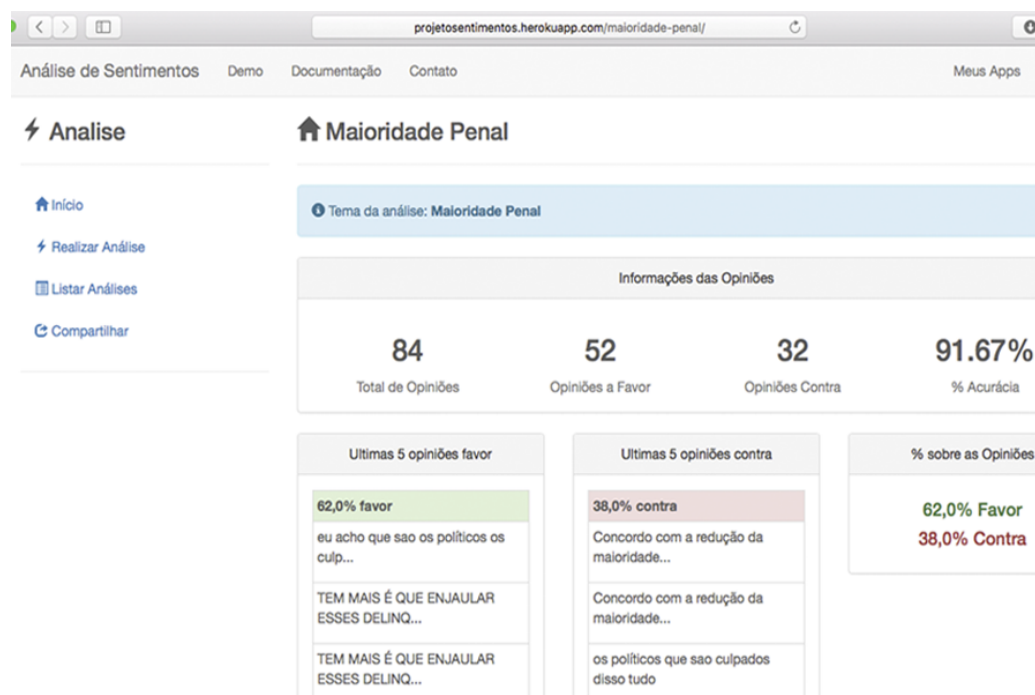


Figura 16 – Tela para verificar acurácia

Fonte: Autor

Listar Análises

Pesquisar

Opinião	Resultado	Correta?
O Adolescente ja tem noção de seus proprios atos	favor	sim
eu acho que tem de diminuir a maioridade e pra 14 anos	favor	sim
rede salesiano manifestacao de repudio a reducao da maioridade penal	contra	sim
tinha que mandar matar esses vagabundos	favor	sim
Leva o champinha p/ sua ksa casa, Dilmaladra terrorista.	favor	sim
O Adolescente ja tem noção de seus proprios atos	favor	sim
sou contra a maioridade penal	contra	sim

Figura 17 – Tela para conferir dados classificados

Fonte: Autor

Apêndice B – Stemmer

Regras de Remoção de Sufixos

Para um melhor entendimento das tabelas de regras é apresentado um exemplo de regra comentado

“inho”, 3, “”, “caminho”, “carinho”, “padrinho”, “sobrinho”, “vizinho”

Quando “inho” é um sufixo que denota um diminutivo, 3 é o tamanho mínimo para o stem, o que evita que palavras como “linho” sofram o stem e palavras entre colchetes são as exceções para esta regra, isto é, elas terminam com o sufixo, mas não só diminutivos. Todas as outras palavras que terminam em –inho e que são mais longas que 6 caracteres sofrerão stemming. Não existe sufixo de reposição para essa regra.

A seguir são mostradas as tabelas contendo as regras para o stemming em Português. A lista de exceções não é apresentada aqui.

Tabela 12 – Regras de redução de plural

Sufixo a Remover	Tamanho Mínimo Stem	Substituição	Exemplo
ns	1	m	bons → bom
ões	3	ão	balões → balão
ães	1	ão	capitães → capitão
ais	1	al	normais → normal
éis	2	el	papéis → papel
eis	2	el	amáveis → amável
óis	2	ol	lençóis → lençol
is	2	il	barris → barril
les	3	l	males → mal
res	3	r	mares → mar
s	2		casas → casa

Fonte: Autor

Tabela 13 – Regras de redução de feminino

Sufixo a Remover	Tamanho Mínimo Stem	Substituição	Exemplo
ona	3	ão	chefona → chefão
ã	2	ão	vilã → vilão
ais	1	al	normais → normal
ora	3	or	professora → professor
na	4	no	amáveis → amável
óis	2	ol	americana → americano
inha	3	inho	chilena → chileno
esa	3	ês	sozinho → sozinha
res	3	r	mares → mar
osa	3	oso	inglesa → inglês
ica	3	ico	maníaca → maníaco
ada	2	ado	prática → prático
ida	3	ido	mantida → mantido
ída	3	imo	prima → primo
iva	3	ivo	passiva → passivo
eira	3	eiro	primeira → primeiro

Fonte: Autor

Tabela 14 – Regras de redução advérbio

Sufixo a Remover	Tamanho Mínimo Stem	Substituição	Exemplo
mente	4		felizmente → feliz

Fonte: Autor

Tabela 15 – Augmentative/diminutive reduction rules

Sufixo a Remover	Tamanho Mínimo Stem	Substituição	Exemplo
díssimo	5		cansadíssimo → cansad
abilíssimo	5		amabilíssimo → ama
íssimo	3		fortíssimo → fort
ésimo	3		
érrimo	4		chiquérrimo → chiqu
zinho	2		pezinho → pe
quinho		c	maluquinho → maluc
uinho	4		amiguinho → amig
adinho	3	r	cansadinho → cansad
inho	3		carrinho → carr
alhão	4		grandalhão → grand
uça	4		dentuça → dent
aço	4		ricaço → ric
adão	4		casadão → cans
ázio	3		corpázio → corp
zão	2		calorzão → calor
ão	3		meninão → menin

Fonte: Autor

Tabela 16 – Regras de redução substantivos

Sufixo a Remover	Tamanho Mínimo Stem	Substituição	Exemplo
encialista	4	c	existencialista → exist
alista	5		minimalista → minim
agem	3		contagem → cont
iamento	4		gerenciamento → gerenc
amento	3		monitoramento → monit
imento	3		nascimento → nas
alizado	4		comercializado → comerci
atizado	4		traumatizado → traum
izado	5		alfabetizado → alfabet
ativo	4		associativo → associ
tivo	4		contraceptivo → contracep
ivo	4		esportivo → esport
ado	2		abalado → abal
ido	3		impedido → imped
ador	3		ralador → ral
edor	3		entendedor → entend
idor	4		cumpridor → cumpr
atória	5		obrigatória → obrig
or	2		produtor → produt
abilidade	5		comparabilidade → compar
icionista	4		abolicionista → abol
cionista	5		intervencionista → interven
ional	4		profissional → profiss
ência	3		referência → refer
ância	4		repugnância → repug
edouro	4		abatedouro → abat
queiro	3		fofoqueiro → fofoq
eiro	3		brasileiro → brasil
oso	3		gostoso → gost
alizaç	4		comercializaç → comerci
ismo	4		consumismo → consum
izaç	5		concretizaç → concret
aç	3		alegaç → aleg
iç	3		aboliç → abol
ário	3		anedotário → anedot
ério	6		ministério → minist
ês	4		chinês → chin
eza	3		beleza → bel
ez	4		rigidez → rigid
esco	4		parentesco → parent
ante	2		ocupante → ocup
ástico	4		bombástico → bomb
ático	3		problemático → problem
ividade	5		produtividade → produt
idade	5		profundidade → profund
oria	4		aposentadoria → aposentad
encial	5		existencial → exist
ista	4		artista → art
quice	4	c	maluquice → maluc
ice	4		chatice → chat

Tabela 17 – Regras para redução de verbos

Sufixo a Remover	Tamanho Mínimo Stem	Substituição	Exemplo
aríamo	2		cantaríamo → cant
ássemo	2		cantássemo → cant
eríamo	2		beberíamo → beb
êssemo	2		bebêssemo → beb
iríamo	3		partiríamo → part
íssemo	3		partíssemo → part
áramo	2		cantáramo → cant
árei	2		cantárei → cant
aremo	2		cantaremo → cant
ariam	2		cantariam → cant
ássei	2		cantássei → cant
assem	2		cantassem → cant
ávamo	2		cantávamo → cant
êramo	3		bebêramo → beb
eremo	3		beberemo → beb
eriam	2		beberiam → beb
eríei	3		beberíei → beb
êssei	3		bebêssei → beb
essem	3		bebessem → beb
íramo	3		partiríamo → part
iremo	3		partiremo → part
iriam	3		partiriam → part
iríei	3		partiríei → part
íssei	3		partíssei → part
issem	3		partissem → part
ando	3		cantando → cant
endo	3		bebendo → beb
indo	3		partindo → part
eria	3		beberia → beb
ermo	3		bebermo → beb
esse	3		bebesse → beb
este	3		bebeste → beb
íamo	3		bebíamo → beb
iram	3		partiram → part
íram	3		concluíram → conclu
irde	3		partirde → part
irei	3		partírei → part
irem	3		partirem → part
iria	3		partiria → part
irmo	3		partirmo → part
isse	3		partisse → part
iste	4		partiste → part
amo	2		cantamo → cant
ara	2		cantara → cant
ará	2		cantara → cant
are	2		cantare → cant
ava	2		cantava → cant
emo	2		cantemo → cant
era	3		bebera → beb
erá	3		beberá → beb

Tabela 18 – Regras de remoção de vogal

Sufixo a Remover	Tamanho Mínimo Stem	Substituição	Exemplo
a	3		menina → menin
e	3		grande → grand
o	3		menino → menin

Fonte: Autor