UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

UMA ARQUITETURA DE PRÉ-PROCESSAMENTO PARA ANÁLISE DE SENTIMENTO EM MÍDIAS SOCIAIS EM PORTUGUÊS BRASILEIRO

DOUGLAS DA ROCHA CIRQUEIRA

DM 27/2018

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém – Pará - Brasil
2018

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

DOUGLAS DA ROCHA CIRQUEIRA

UMA ARQUITETURA DE PRÉ-PROCESSAMENTO PARA ANÁLISE DE SENTIMENTO EM MÍDIAS SOCIAIS EM PORTUGUÊS BRASILEIRO

DM 27/2018

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém – Pará - Brasil
2018

UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

DOUGLAS DA ROCHA CIRQUEIRA

UMA ARQUITETURA DE PRÉ-PROCESSAMENTO PARA ANÁLISE DE SENTIMENTO EM MÍDIAS SOCIAIS EM PORTUGUÊS BRASILEIRO

Dissertação submetida à banca examinadora do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará para a obtenção do Grau de Mestre em Engenharia Elétrica na área de Computação Aplicada.

UFPA / ITEC / PPGEE

Campus Universitário do Guamá

Belém – Pará - Brasil

2018

UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

UMA ARQUITETURA DE PRÉ-PROCESSAMENTO PARA ANÁLISE DE SENTIMENTO EM MÍDIAS SOCIAIS EM PORTUGUÊS BRASILEIRO

AUTOR: DOUGLAS DA ROCHA CIRQUEIRA

DISSERTAÇÃO DE MESTRADO SUBMETIDA À AVALIAÇÃO DA BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DA UNIVERSIDADE FEDERAL DO PARÁ E JULGADA ADEQUADA PARA OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM ____/____/____

BANCA EXAMINADORA:

_____

Prof. Dr. Ádamo Lima de Santana
ORIENTADOR - UFPA

_____

Prof. Dr. Marcelino Silva da Silva
MEMBRO INTERNO – PPGEE

_____

Prof. Dr. Marcos Cesar da Rocha Seruffo
MEMBRO EXTERNO – FCT-UFPA

VISTO:

_____

Profa. Dra. Maria Emília de Lima Tostes

COORDENADORA DO PPGEE/ITEC/UFPA

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter permitido que eu entrasse em uma universidade e concluísse minha graduação, para então chegar ao Mestrado.

Agradeço a minha mãe, Maria Elis da Rocha Cirqueira, que sempre se esforçou para me educar e para tornar-me a pessoa que sou hoje. Agradeço ao meu pai, Artur Alves Cirqueira, o grande exemplo de superação, força, bondade e humildade que tenho em minha vida.

Agradeço a minha esposa, Milly Juliana Pinto Corrêa Cirqueira, que esteve comigo em todos os momentos difíceis dessa caminhada, e se mostrou a melhor companheira que eu poderia sonhar em ter na vida.

Agradeço a minha família, que sempre me apoiou e torceu pelo meu sucesso.

Agradeço aos meus amigos, que sempre ajudaram com conselhos e discussões. Em especial ao Leandro Fonseca Chaves, Cláudio Rogério, Ronedo de Sá Ferreira, Filipe Vieira e Prof. Dr. Eduardo Cerqueira, que contribuíram ao longo da minha caminhada desde a graduação.

Agradeço ao meu orientador, Prof. Dr. Ádamo Lima de Santana, por me aceitar em seu laboratório e ter permitido que eu seguisse o caminho profissional que escolhi para a minha vida.

Agradeço ao Laboratório de Inteligência Computacional, em especial a Márcia Fontes Pinheiro, Prof. Fábio Lobato e Prof. Antônio Jacob, por todo o apoio e direcionamentos em projetos desenvolvidos no contexto do laboratório.

Agradeço aos gerentes Igor Araújo, Lucas Vinicius, aos colegas de laboratório Gean Costa, Vincent Tadaiesky, Abner Cardoso, Beatriz Nery, Douglas Alvarez, Arthur Cordovil, Felipe Cardoso-Shounen, Lucas Novais e Marcus Carvalho, e a todos os demais membros da família LINC, por todas as discussões e apoio em atividades desenvolvidas no laboratório.

Agradeço à Universidade Federal do Pará por todas as oportunidades e ensinamentos. Um ambiente onde aprendi a lidar com diversidades, adversidades e superar desafios. Agradeço também por todo apoio dado para projetos e trabalhos desenvolvidos ao longo do Mestrado. Agradeço a CAPES pelo fundamental fomento ao desenvolvimento desta dissertação e projeto.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

AS	–	Análise de Sentimento
PT_Br	–	Português Brasileiro
API	–	*Application Development Interface*
CSV	–	*Comma-Separated Values*
MO	–	Mineração de Opinião
MT	–	Mineração de Texto
IE	–	*Information Extraction*
IR	–	*Information Retrieval*
KDD	–	Knowledge Discovery in Databases
NLP	–	*Natural Language Processing*
UGC	–	User-Generated Content ou Conteúdo Gerado por Usuário
SVM	–	Support Vector Machine
RSO	–	Redes Sociais Online
TDM	–	*Text Data Mining*
KDT	–	*Knowledge Discovery from Textual Databases*
ML	–	*Machine Learning*
TP	–	*Term Presence*
TF	–	*Term Frequency*
TF-IDF	–	*Term Frequency – Inverse Document Frequency*

RESUMO

A Web 2.0 e a evolução nas Tecnologias da Informação e Comunicação, têm impulsionado novos meios de interação e relacionamento. Neste contexto, as Redes Sociais Online (RSO) são um exemplo, como plataformas que permitem a interação e o compartilhamento de informações entre pessoas. Além disso, é possível observar que RSO passaram a ser adotadas como canal de desabafo de consumidores, por meio de opiniões sobre produtos e experiências. Este cenário apresenta uma ótima oportunidade para que empresas possam melhorar produtos, serviços e estratégias de mercado, já que as RSO são poderosas fontes massivas de dados não-estruturados gerados pelo consumidor (do inglês, *User-Generated Content* - UGC), com opiniões e avaliações sobre ofertas em plataformas tais como Facebook, Twitter e Instagram. O Brasil é um grande exemplo onde esse fenômeno pode ser observado e apresenta potencial oportunidade de exploração de mercado, dado que a população brasileira é uma das nações que mais utiliza RSO no mundo. Neste âmbito, técnicas computacionais de Mineração de Opinião (MO) ou Análise de Sentimento (AS) são aplicadas com o intuito de inferir a polaridade dominante (positivo, negativo, neutro) quanto ao sentimento associado a textos, e, podem ser aplicadas em dados de RSO a fim de avaliar o feedback do público-alvo. Apesar das diversas estratégias de AS reportadas na literatura, ainda há vários desafios enfrentados na aplicação de AS em textos oriundos de RSO, devido às características da linguagem utilizada em tais plataformas. O estado da arte de AS é voltado para a língua inglesa e as propostas existentes para Português Brasileiro (PT_Br) não apresentam uma metodologia padronizada nas tarefas de pré-processamento. Neste âmbito, esta pesquisa investiga uma metodologia sem tradução e propõe uma nova arquitetura expandida de pré-processamento de AS voltada para o PT_Br, a fim de prover atributos enriquecidos para os algoritmos de AS. A proposta foi comparada com modelos bem estabelecidos na literatura, e resultados obtidos indicam que esta pode superar o estado da arte em até 3% de revocação, para 6 de 7 bases de dados avaliadas.

PALAVRAS-CHAVE: Análise de Sentimento; Pré-processamento; Processamento de Linguagem Natural; Mineração de Texto; Mineração de Opinião; Mineração de Dados; Redes Sociais Online; Mídias Sociais

ABSTRACT

The Web 2.0 and the evolution of Information Technologies have brought novel interaction and relationship channels. In this context, the Online Social Networks (OSN) are an example as platforms which allow interactions and sharing of information between people. In this scenario, it is possible to observe the adoption of OSN as a channel for posting opinions regarding products and experience. This scene presents an excellent opportunity for companies that aim to improve products, services and marketing strategies, given OSNs are powerful sources of massive unstructured data generated by consumers (UGC), with opinions and reviews concerning offers, in platforms such as Facebook, Twitter and Instagram. Brazil is a highlight in this scenario, where this phenomenon can be observed, as the Brazilian population is one of the most active in social media platforms in the world. This makes it a country full of opportunities to market exploitation. In this context, computational techniques of Opinion Mining and Sentiment Analysis (SA) are applied aiming to infer the polarity (positive, negative, neutral) regarding a sentiment associated to texts, and can also be applied in data from OSN to evaluate the feedback from a target audience. Although the existing diversity of SA strategies reported in the literature, there are still challenges faced in the application of SA in text data from OSN, given the characteristics of the language adopted in such platforms. The state of art is focused on SA towards the English language, and the existing proposals for Brazilian Portuguese do not have a standardized methodology for preprocessing steps. In this context, this research investigates an approach with no translation, and proposes a novel preprocessing architecture for SA towards Brazilian Portuguese, aiming to provide enriched features to SA algorithms. The proposal was compared with well-established baselines from the literature, and the obtained results indicate that this architecture can overcome the state of art recall in at least 3% , for 6 out of 7 datasets evaluated.

KEYWORDS: Sentiment Analysis; Preprocessing; Natural Language Processing; Text Mining; Opinion Mining; Data Mining; Social Networks; Social Media

1 INTRODUÇÃO

1.1 DEFINIÇÃO DO PROBLEMA E MOTIVAÇÃO

Com o advento das Tecnologias da Informação e Comunicação, Internet e Multimídia, diversas ferramentas e plataformas para a Web 2.0 emergiram e transformaram os meios de comunicação, interação, relacionamento e aquisição de conhecimento (VENKATESH, CROTEAU e RABAH, 2014; COOK, 2017).

Neste contexto, surgiram as Redes Sociais Online (RSO) que são plataformas que permitem a interação entre pessoas, o compartilhamento de informações e a composição de grupos em ambientes online (SANTANA, V.; *et al.*, 2009). Com essas plataformas, os usuários têm acesso às ferramentas para interação dinâmica com outros usuários de forma síncrona e assíncrona.

Neste cenário, também é possível observar a tendência de utilizar as RSO não somente para entretenimento, mas para compartilhamento de opiniões em relação a produtos, serviços, marcas e experiências (FAN; GORDON, 2014). Plataformas com propósitos específicos e funcionalidades de RSO tem surgido, tais como Facebook[1], Twitter[2] e Instagram[3], e que apresentam milhões de usuários ativos que postam os mais variados tipos de conteúdo, seja ele textual, visual ou em áudio.

As RSO, do ponto de vista dos consumidores, representam canais para coletar informações sobre marcas, produtos e serviços, assim como para compartilhar opiniões e experiências, também chamados de conteúdo gerado pelo consumidor ou UGC (YU; DUAN; CAO, 2013). Da perspectiva das empresas, as RSO são poderosas fontes de dados sobre o comportamento dos consumidores, além de um novo canal de comunicação e de baixo custo (LOBATO, *et al*., 2016).

As comunidades da academia e indústria de diversas áreas, tais como marketing, comunicação, ciência política, entre outras, já perceberam a importância do UGC oriundo das RSO, já que produtos, serviços e estratégias de negócios podem ser adaptados de acordo com o feedback do público-alvo (LEGGAT, 2011). Dessa forma, é possível

[1] https://www.facebook.com/
[2] https://twitter.com/
[3] https://www.instagram.com/

oferecer serviços, produtos ou atendimento como o cliente deseja ou espera das instituições, e assim conseguir alcançar uma fatia maior de mercado com um número maior de clientes e um bom posicionamento de marca (GARRIGOS, *et al*., 2012).

Nesse cenário, o Brasil é o segundo país no *ranking* de população adepta de redes sociais no mundo (WE ARE SOCIAL & HOOTSUITE, 2018), sendo considerado a capital do universo em mídias sociais (CHAO, 2013), em razão de os brasileiros serem muito comunicativos e gastarem muito tempo utilizando tais plataformas para compartilhamento de mídias e informações (WE ARE SOCIAL & HOOTSUITE, 2018).

Em 2016, segundo pesquisa do Instituto Brasileiro de Geografia e Estatística – IBGE (2018), o Brasil possuía 116 milhões de usuários de Internet. O relatório divulgado pela Conferência das Nações Unidas sobre o Comércio e Desenvolvimento (*United Nations Conference on Trade And Development* – UNCTAD, 2017) corrobora com esses indicadores, apontando que o Brasil ocupa a quarta posição na lista de países com mais pessoas conectadas à Internet, com 120 milhões de usuários.

Também em 2016, o Comitê Gestor da Internet no Brasil (CGI.br), por meio do Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (CETIC.br) do Núcleo de Informação e Coordenação do Ponto BR (NIC.br)(2016), apresentou em pesquisa que 78% dos internautas brasileiros possuíam acesso a pelo menos uma RSO. Segundo estatísticas do Facebook, 102 milhões de brasileiros têm contas ativas na plataforma[4].

Pesquisa da Social@Ogilvy e SurveyMankey [5] mostrou que 94% dos consumidores brasileiros seguem alguma marca nas RSO, 81% destes seguem para mostrar satisfação com a marca ou produto e porque as empresas normalmente respondem seus comentários nessas plataformas. Além disso, 95% dos brasileiros que participaram da pesquisa afirmaram que gostam de compartilhar boas experiências com marcas ou produtos nas RSO, enquanto que 78% dos brasileiros afirmam também utilizar as RSO para compartilhar experiências ruins com essas companhias.

[4] 102 milhões de brasileiros compartilham seus momentos no Facebook todos os meses. https://www.facebook.com/business/news/102-milhes-de-brasileiros-compartilham-seus-momentos-no-facebook-todos-os-meses

[5] Social@Ogilvy & SurveyMonkey. How to cultivate brand advocacy. Disponível em <https://pt.slideshare.net/socialogilvy/how-to-cultivate-brand-advocacy>

A pesquisa também revelou que 42% dos brasileiros entrevistados consideram-se promotores de marcas nas RSO, recomendando produtos e serviços para amigos e familiares. Destes, 37% acreditam que as recomendações de um amigo próximo através de uma RSO são mais confiáveis do que recomendações de conhecidos nas mesmas plataformas (15%) e recomendações feitas pessoalmente (22%). Os outros 26% dos brasileiros acreditam que os três tipos de recomendações supracitadas são igualmente confiáveis.

O estudo da PricewaterhouseCoopers (2016) ratifica essa confiança dos brasileiros nas mídias sociais, apresentando que 77% dos mesmos são influenciados na decisão de compra através das informações obtidas nas RSO pelos comentários e curtidas de amigos ou perfis de marcas.

Percebendo essa oportunidade de exploração de potencial mercado, muitas empresas brasileiras já se convenceram sobre a importância do uso de RSOs e agregar as informações oriundas destas plataformas, transformando-as em inteligência de negócio, valor para à empresa e resultando numa experiência superior para o cliente. De acordo com pesquisa do CGI.br, por meio do CETIC.br do NIC.br (2018) sobre a presença das TICs em microempresas, 70% das microempresas e 71% das grandes empresas brasileiras possuem perfis em plataformas de RSO.

No entanto, segundo estudo da Sprinklr (2016) as empresas brasileiras ainda não conseguem integrar as informações dos consumidores que são adquiridas das RSO com as demais estratégias da empresa. Embora o monitoramento e análise das informações das mídias sociais possa melhorar os resultados e *insights* sobre o consumidor, isso ainda é uma tarefa difícil de ser realizada devido ao volume de opiniões textuais e às características não estruturadas e esparsas dos dados gerados nestas plataformas. Esta tarefa é impossível de ser realizada manualmente por humanos (LIU & ZHANG, 2012).

A fim de lidar com essa problemática, técnicas de Processamento de Linguagem Natural (do inglês, *Natural Language Processing* - NLP)(GRISHMAN, 1984; CHOWDHURY, 2003), Extração da Informação (do inglês, *Information Extraction* - IE), Recuperação de Informação (do inglês, *Information Retrieval* - IR) (SALTON & MCGILL, 1986), Descoberta de conhecimento de Conhecimento em Base de Dados (do inglês, Knowledge Discovery in Databases – KDD)(FAYYAD, *et al*. 1996), Mineração da Web (COOLEY, MOBASHER e SRIVASTAVA, 1997) e Mineração de Texto

(MT)(FELDMAN & DAGAN, 1995; FELDMAN & SANGER, 2007), têm sido aplicadas para obter *insights* a partir dos dados das RSO. Além disso, o campo de estudo da Análise de Sentimento ou Mineração de Opinião busca detectar, extrair e classificar opiniões, sentimentos, emoções, avaliações e atitudes relativas a diferentes tópicos ou entidades, como produtos, serviços, organizações, indivíduos e eventos (PANG, et al., 2008; Liu, 2012).

Diversas estratégias já foram propostas para realizar AS nos mais variados tipos de texto, que podem variar quanto ao seu gênero, estilo de escrita e origem. Muitos desses trabalhos foram desenvolvidos para linguagens específicas e, dentre elas, o inglês é encontrado com maior frequência.

Embora alguns trabalhos para AS também já foram propostos especificamente para o PT_Br (SOUZA, ELLEN, *et al*., 2016), estes geralmente trabalham apenas com *reviews* relacionadas a algum tipo de produto ou dados de redes sociais, tal como o Twitter (SANTOS, CAROLINE, *et al*, 2016). Entretanto, como o estado da arte dos métodos de AS é voltado para a língua inglesa (PIRYANI; MADHAVI; SINGH, 2017), quando se tenta adotar o processo de tradução automática de textos escritos em outras línguas para aplicar a AS em português, o resultado por vezes não é satisfatório (CIRQUEIRA, et al. 2016).

Assim, esta pesquisa teve início a partir de revisões de literatura em AS. Dentre as lacunas encontradas na literatura, observou-se a ausência de bases de dados públicas para replicação de experimentos Souza, *et al*. (2016a, 2016b), AS multilíngue sem tradução automática e bases de dados mais ferramentas para outras línguas que não o inglês Su, Luo e Chen (2016).

Consequentemente, viu-se nestas lacunas uma motivação para revisão da literatura e busca por problemáticas em AS no contexto universal e brasileiro para mídias sociais. Nesta ótica, foram realizadas análises de publicações mais recentes, além de teses e dissertações. Porém, alguns trabalhos antigos se fazem notar devido à relevância para o tema.

Foram observados diversos desafios presentes no cenário UGC, incluindo informalidades e características intrínsecas da linguagem adotada na Internet. Ainda, notou-se uma problemática comum na área de AS, a qual se caracteriza como a ausência

de padronização em uma das etapas desta técnica, a de pré-processamento Souza, *et al*. (2016a, 2016b). Esta etapa visa o tratamento dos dados para a aplicação da técnica de AS.

Notou-se que existem diversas tarefas de pré-processamento reportadas na literatura, que lidam com as particularidades da linguagem textual a ser dada como entrada para a AS. Algumas destas tarefas visam capturar diretamente sinais de sentimento, e outras são mais focadas na limpeza dos dados de texto.

Neste sentido, notou-se que cada proposta apresenta suas próprias tarefas de pré-processamento. Todavia, não foi observado um estudo que expandisse tais estágios, a fim de cobrir o máximo possível das estratégias mencionadas na literatura, focada no cenário de RSO para PT_BR.

Como contribuição deste estudo tem-se então a proposta de uma arquitetura especializada de pré-processamento de dados escritos em PT_BR e oriundos de RSO para AS, denominada SentiBR. Esta será referenciada por vezes ao longo desta dissertação como arquitetura ou *framework.*

Esta proposta tem motivação na falta de padronização no processo de pré-processamento de dados, o qual permeia as publicações deste campo para PT_BR. Isto pode representar um empecilho para a plena utilização dos métodos de AS propostos pela comunidade científica na academia.

O caráter inovador desta arquitetura se faz presente no fato de que a mesma é a que incorpora o maior número de tarefas de pré-processamento para AS voltada ao PT_Br em dados de RSO. Além deste elemento, foi adotada uma metodologia de implementação inédita para algumas tarefas de pré-processamento, o que nunca havia sido testado anteriormente no cenário brasileiro de AS.

A fim de comprovar a viabilidade da metodologia proposta, o SentiBR foi comparado com sólidos modelos existentes na literatura, almejando uma comparação com o estado da arte. Os resultados obtidos através de estudos de casos em bases representativas de mídias sociais indicam que um conjunto expandido de pré-processamento, pode melhorar o desempenho de AS no contexto do PT_Br em até 3% na identificação de verdadeiros positivos (revocação), e enriquecer as possíveis análises referentes às opiniões de consumidores e usuários em RSO. Além disto, este modelo superou o estado da arte em 6 das 7 bases de dados avaliadas, sendo a base em que não

se obteve o maior desempenho composta de notícias, o que não é o cenário foco do SentiBR.

Por fim, para a condução da pesquisa definiu-se as seguintes hipóteses e objetivos, conforme descrito a seguir.

1.2 HIPÓTESES E OBJETIVOS

Considerando o problema exposto e as motivações deste trabalho, definiu-se a seguinte hipótese que rege esta dissertação:

- Uma arquitetura expandida de pré-processamento de dados para conteúdo gerado por consumidores usuários em Redes Sociais Online, pode melhorar o desempenho de tarefas de Mineração de Texto neste domínio, tal como a Análise de Sentimento.

Com o intuito de avaliar a hipótese, definiu-se os seguintes objetivos:

1. Identificar as tarefas de pré-processamento que possibilitem o tratamento de dados oriundos de RSO para aplicar a técnica de AS;

2. Aplicar e analisar a metodologia em experimentos com dados e *baselines* do estado da arte, verificando a viabilidade de adoção da arquitetura desenvolvida através de cenários de teste com características contextuais influenciadoras de desempenho, no contexto de mídias sociais.

O primeiro objetivo visa abalizar o processo de desenvolvimento da metodologia, o qual deverá ser dividido em tarefas sequenciais que irão enriquecer o conjunto de dados para então aplicar a AS em dados de RSO.

O segundo objetivo visa aplicar e analisar a viabilidade da metodologia proposta com dados reais de RSO, de maneira que indicadores de desempenho possam ser mensurados de acordo com os desafios presentes neste contexto.

A avaliação de desempenho é realizada a partir de sete bases de dados disponibilizadas por trabalhos presentes na literatura brasileira de AS. A avaliação de desempenho possui quatro etapas:

1. Testar a arquitetura proposta com um algoritmo de aprendizado de máquina em todas as bases de dados dispostas, para elencar um panorama acerca do desempenho geral comparado às demais metodologias no estado da arte, a fim de se investigar a resposta para a hipótese levantada;
2. Fazer levantamento estatístico dos desempenhos da arquitetura proposta e dos *baselines* do estado da arte nos experimentos desenvolvidos;
3. Analisar o desempenho da proposta deste estudo sob a ótica do *trade-off* entre a precisão dos resultados e tempo de execução;
4. Identificar o desempenho de cada tarefa de pré-processamento quando empregada individualmente;

Com isso, espera-se possibilitar uma padronização na metodologia de pré-processamento em AS para o PT_Br, de modo a impulsionar os estudos relacionados. A metodologia proposta emprega um número expandido de tarefas de tratamento de dados reportadas no estado da arte de AS e foi comparada com sólidos modelos existentes na literatura através da utilização do algoritmo *Support Vector Machine* (SVM) para classificação do sentimento, almejando uma comparação com o estado da arte.

Os resultados obtidos através de estudos de casos em bases reportadas na literatura indicam uma melhora na AS no contexto do PT_Br. Outro resultado encontrado foi referente a novos indicadores de desempenho quanto a diferentes tarefas de pré-processamento de dados. Estes podem prover aos pesquisadores nesta área, informações acerca das melhores estratégias para se trabalhar com AS para Mineração de Opinião de consumidores em plataformas da Web 2.0.

1.3 ORGANIZAÇÃO DO DOCUMENTO

O presente trabalho está organizado da seguinte forma: O Capítulo 2 aborda os principais fundamentos teóricos sobre Análise de Sentimento, bem como sobre o estilo de comunicação e linguagem encontrada na internet, principalmente em Redes Sociais. No Capítulo 3 apresenta-se trabalhos correlatos ao tema analisado, enfatizando o estado da arte quanto à Análise de Sentimento em dados de Redes Sociais Online. O Capítulo 4 explica os passos desenvolvidos para realizar a implementação da proposta deste projeto.

O Capítulo 5 apresenta os testes e resultados alcançados, e o Capítulo 6 apresenta as conclusões referentes aos resultados deste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 CONSIDERAÇÕES INICIAIS

A convergência da computação e comunicação dinâmica nas plataformas presentes na *Web* 2.0, têm produzido uma sociedade que é consumidora e produtora de informação. A agência de marketing Zephoria[6] revela que cerca de 300 milhões de fotos são postadas por dia no Facebook, enquanto que a plataforma Internet Live Stats[7] reporta que 350 mil *tweets* são postados a cada minuto. Neste contexto, faz-se necessário o desenvolvimento de técnicas para analisar este *Big Data*, a fim de se explorar potenciais *insights* para organizações privadas e públicas.

No campo da Mineração de Dados, existe uma área específica com ênfase em Mineração de Texto. Esta é composta de técnicas com ênfase no Processamento de Linguagem Natural, o qual também engloba o campo de AS. Através desta, é possível a mineração de reclamações e elogios, que podem ser úteis para companhias ao redor do mundo.

Neste âmbito, a MT e AS podem ser e têm sido aplicadas em diversas áreas do conhecimento para os mais diversos domínios como Bioinformática (PLETSCHER-FRANKILD; *et al.*, 2015)(DAI; *et al.*, 2015)(CEJUELA; *et al.,* 2017)(BADAL; KUNDROTAS; VAKSER, 2018), Medicina (KUSHIMA; *et al.*, 2017)(BEJAN; *et al*, 2017)(CHEN; BARBOUR, 2017), Segurança da Informação (LING*, et al*., 2015), Mercado Financeiro (SUN; LACHANSKI; FABOZZI, 2016)(LI, 2017)(KIM; *et al*. 2017)(SEO; PARK, 2018)(ELAGAMY, STAINER; SHARP, 2018), Inteligência de Negócios (BEREZINA; et al., 2016)(MORO; CORTEZ; RITA, 2015) e Gestão de Relacionamento com o Cliente utilizando mídias sociais (do inglês, *Social Customer Relationship Management* – Social CRM)(YEE LIAU; PEITAN, 2014)(BUETTNER, 2017)(ADAMOPOULOS; GHOSE; TODRI, 2018).

[6] ZEPHORIA DIGITAL MARKETING. The Top 20 Valuable Facebook Statistics. Disponível em <https://zephoria.com/top-15-valuable-facebook-statistics/>

[7] INTERNET LIVE STATS. Twitter Usage Statistics. Disponível em <http://www.internetlivestats.com/twitter-statistics/>

Algumas técnicas de AS trabalham com listas de palavras afetivas e sua detecção em documentos de texto, enquanto que outros métodos operam através de algoritmos sofisticados de aprendizado de máquina.

Assim, este capítulo apresentará a fundamentação teórica sobre Mineração de Texto, destacando as principais etapas do processo, seguindo para os fundamentos sobre AS, apresentando um breve histórico, estado da arte, os principais algoritmos utilizados, e as métricas de avaliação de resultados.

Além disso, este capítulo também aborda detalhes quanto aos conceitos das plataformas de mídia e rede social, e suas particularidades. Serão também fornecidos detalhes quanto às duas plataformas de RSO com as quais se realizam os experimentos desta dissertação, que são o Twitter e Facebook. Por fim, particularidades da linguagem adotada em RSO serão ilustradas.

Os conceitos vistos neste capítulo servem como embasamento teórico para a metodologia e arquitetura proposta nesta dissertação, dado que os experimentos realizados trabalham com dados textuais provenientes de plataformas de RSO. Tais aspectos serão mais detalhados também nos Capítulos 3 e 4.

2.2 MINERAÇÃO DE TEXTO

A Mineração de Texto, também chamada de Mineração de Dados Textuais (*Text Data Mining* - TDM)(FELDMAN; DAGAN, 1995) ou Descoberta de Conhecimento em Bases Textuais (*Knowledge Discovery from Textual Databases* - KDT)(HEARST, 1997) consiste de uma forma geral, em um processo de conhecimento intensivo no qual um analista de dados interage com uma coleção de documentos ao longo do tempo, usando um conjunto de ferramentas de análise.

Analogamente ao processo de KDD, a MT busca extrair informações úteis de fontes de dados através da identificação e exploração de padrões interessantes. Porém, no caso da MT, as bases de dados são coleções de documentos, também chamados de *corpus.* Nestes, os padrões não são encontrados em registros de bases de dados estruturadas, mas sim em dados textuais não estruturados nos documentos dessas coleções (FELDMAN; SANGER, 2007).

A MT é um campo interdisciplinar que tem como base técnicas de Processamento de Linguagem Natural (GRISHMAN, 1984; CHOWDHURY, 2003), Extração da Informação e Recuperação de Informação (SALTON & MCGILL, 1986) e KDD (FAYYAD, *et al*. 1996). Assim, a MT pode ser vista como uma aplicação de técnicas de KDD sobre dados textuais.

Os autores (FELDMAN; SANGER, 2007) acrescentam que a Mineração de Texto utiliza um conjunto de técnicas KDD, Aprendizado de Máquina ou *Machine Learning* (ML), NLP, IR e Gestão de Conhecimento. Estas envolvem o processamento de coleções de documentos, armazenamento de representações intermediárias, técnicas (análise estatística, classificação, clusterização, regras de associação, etc.) para analisar tais informações e visualização dos resultados. Ademais, Feldman e Dangan (1995) afirmam que antes de aplicar qualquer técnica de KDD em dados textuais, é necessário extrair representações estruturadas destes.

Nas seções a seguir são apresentadas as etapas do processo de Mineração de Texto necessárias a realização de extração de conhecimento a partir de dados textuais. Estes conceitos são importantes e embasam o *pipeline* de AS aplicado a dados de RSO.

2.2.1 Etapas do Processo de MT

O processo de MT pode ser considerado uma especialização do processo de KDD. A Figura 2.1 ilustra o processo de MT o qual é basicamente composto por um conjunto de atividades contínuas, interativas e iterativas: Identificação do problema, coleta de dados, seleção de dados, pré-processamento, formatação, Mineração de Dados (MD) e Interpretação e avaliação dos resultados.

2.2.1.1 Definição do Problema

A primeira etapa do processo de MT é a compreensão do domínio de aplicação e a identificação dos objetivos da aplicação do processo de MT, bem como a definição das perguntas de pesquisa.

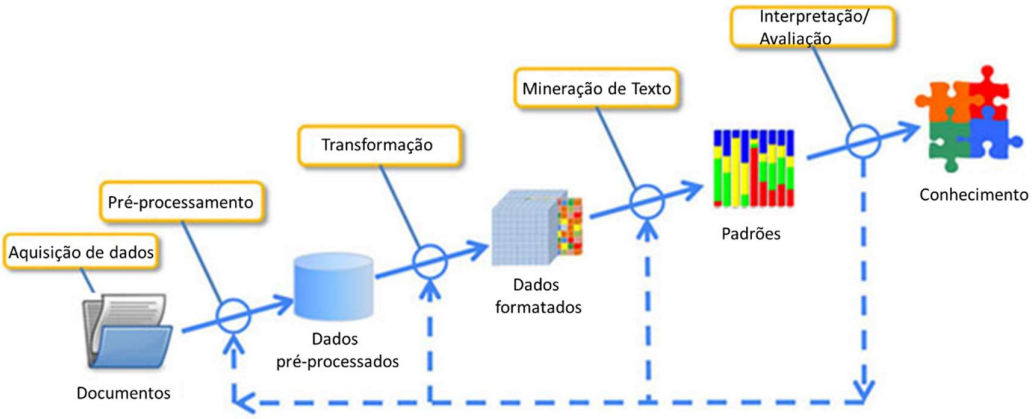

Figura 2.1 Etapas do processo de KDD. Figura adaptada de T-Lab[8].

2.2.1.2 Aquisição de Dados

A segunda etapa do processo de MT é a aquisição dos dados textuais que serão processados. Nesta etapa é feita a coleta de um conjunto de dados de interesse ou foco em um subconjunto, criando assim o corpus no qual o processo de MT será aplicado.

Normalmente o *corpus* é formado por uma coleção de documentos, os quais consistem como unidade de dados textuais que geralmente, mas não necessariamente, correlacionam com algum documento real como relatórios de negócios, memorandos, e-mails, pesquisas, manuscritos, artigos, comunicados de imprensa, histórias, notícias, postagens em RSO, etc. (FELDMAN; SANGER, 2007). Ressalta-se que o *corpus* tem como características um conjunto de dados textuais escritos em linguagem natural, que é inerentemente não estruturado e *fuzzy* (TAN, 1999).

Após coletar o *corpus*, o processo de MT passa a etapa de pré-processamento de dados.

2.2.1.3 Pré-processamento de Dados

A terceira etapa do processo de MT é o processo de limpeza e pré-processamento de dados. Esta etapa é uma das mais importantes em MT, pois engloba diversas tarefas de NLP responsáveis por normalizar e transformar os dados textuais escritos em

[8] https://tlab.it/en/presentation.php

linguagem natural e não estruturados do *corpus* em representações intermediárias explicitamente estruturadas (FELDMAN; SANGER, 2007).

Dado a enorme quantidade de palavras, frases e sentenças que um documento pode ter, uma tarefa essencial para os sistemas de MT é a identificação de um subconjunto de características que possam ser utilizadas para representar este documento como um todo.

Via de regra, as tarefas de MT apresentam elevado custo computacional para processamento, devido à alta dimensionalidade dos dados textuais, pois cada característica representativa do documento pode ser vista como uma dimensão. Este aspecto é um fator determinante no desenvolvimento das operações de pré-processamento de MT, visando assim a criação de modelos de representação mais simplificados. Dependendo da representação utilizada, pode surgir o aspecto de esparsidade na representação dos documentos escritos em linguagem natural.

Dessa forma, há um *trade-off* entre manter uma representação que preserve o significado semântico, as relações entre os objetos do documento e a identificação de características representativas de uma maneira computacionalmente mais eficiente e prática para a descoberta de padrões.

A fim de normalizar, padronizar e selecionar as melhores características de uma base de dados textual, diversas tarefas de NLP são reportadas na literatura e podem ser aplicadas antes da transformação dos dados textuais. As subseções a seguir apresentarão as principais tarefas de pré-processamento de texto oriundas de NLP e utilizadas na literatura.

2.2.1.3.1 Tokenization

Na tarefa de tokenização ou *tokenization*, o texto é dividido em *tokens*, que são segmentos de texto com um significado (e.g. caracteres, palavras, frases) através de modelos de expressões regulares que eliminam caracteres especiais e retornam os *tokens* separados.

2.2.1.3.2 Minúsculas

Na tarefa de Minúsculas é realizada a conversão do texto para letras minúsculas. Esta tarefa permite executar a normalização dos dados para que termos com a mesma grafia, porém que estejam com capitalização distinta, possam ser reconhecidos como o mesmo termo. Por exemplo, os termos “vida” e “VIDA”, serão considerados como *tokens* diferentes se não houver a normalização ou conversão para letras minúsculas.

2.2.1.3.3 Remoção de Acentuação

Esta faz-se necessária em línguas que utilização acentuação, como o PT_BR. O objetivo é normalizar os dados para evitar o problema de omissão de acentos em palavras do *corpus*, seja por erro de digitação ou por estarem escritos na forma informal da língua.

2.2.1.3.4 Tratamento de Pontuação

A tarefa de tratamento de pontuação visa inserir corretamente sinais de pontuação, a fim de melhorar a sintaxe e separação em *tokens*.

2.2.1.3.5 Remoção de Pontuação

Esta tarefa visa a remoção de pontuação do *corpus,* a fim de reduzir o número de *tokens* para representação.

2.2.1.3.6 Remoção de *Stopwords*

A remoção de *stopwords* consiste na remoção de algumas palavras extremamente comuns que parecem ter pouco valor na representação de informação dos documentos. Estas palavras são chamadas de *stopwords*. A estratégia geral para determinar uma lista de *stopwords* é classificar os termos de acordo com a frequência da coleção (i.e. o número total de vezes que cada termo aparece na coleção) e depois são escolhidos os termos mais frequentes. Muitas vezes são também filtradas manualmente pelo conteúdo semântico relativo ao domínio dos documentos. Os membros desta lista de *stopwords* normalmente são removidos durante a identificação de *tokens* nos documentos.

2.2.1.3.7 Tratamento de Abreviações

Esta tarefa visa substituir as abreviações por seus correspondentes expandidos. Por exemplo, o termo “tdb”, quando encontrado em um texto, pode ser substituído por “tudo de bom”.

2.2.1.3.8 Tratamento de Numerais

A tarefa de tratamento de numerais visa remover ocorrências de números no texto ou de convertê-los para a sua forma extensa escrita em texto.

2.2.1.3.9 Correção Ortográfica

Esta tarefa visa aplicar a correção ortográfica em erros encontrados em um texto. Os termos escritos erroneamente são substituídos pela sua forma correta.

2.2.1.3.10 Correção Ortográfica Fonética

Esta tarefa visa a correção ortográfica no texto, mas os erros tratados têm motivação relacionada à fonética da língua. Seu tratamento também se dá pela substituição de erros por versões corretas no texto.

2.2.1.3.11 Stemming

Considerando que os textos escritos em linguagem natural possuem aspectos morfológicos como temas, afixos e desinências que produzem palavras distintas, o *corpus* proveniente de uma coleção de documentos geralmente é muito grande devido ao número de palavras distintas. Por exemplo, palavras com o mesmo significado, como “trabalhar”, “trabalha” e “trabalhando” podem ser separadas como *tokens* diferentes.

A fim de diminuir o número de diferentes *tokens* em um corpus e aumentar a frequência de ocorrência de outros *tokens*, é necessário converter os *tokens* para uma forma padrão. Este processo é chamado *Stemming*. O mesmo reduz palavras flexionadas para sua raiz (HU; LIU, 2012). Por exemplo, “trabalhar”, “trabalha”, “trabalhará”, “trabalhei” e “trabalhando” podem ser representadas pelo radical “trabalh” como *token*.

Os algoritmos de *Stemming* são dependentes das regras da língua à qual estão sendo aplicados.

2.2.1.3.12 Part-Of-Speech Tagging (POS)

A marcação POS realiza a anotação de palavras com *tags* de acordo com o contexto em que estas aparecem. As *tags* dividem as palavras em categorias com base na função que desempenham na sentença em que aparecem e as mesmas fornecem informações sobre o conteúdo semântico de uma palavra. O conjunto mais comum de *tags* é composto por artigo, substantivo, verbo, adjetivo, preposição, número e nome próprio.

2.2.1.3.13 Lemmatization

O processo de lematização ou *lemmatization* consiste na análise morfológica de palavras, com o intuito de remover terminações flexionais e identificar a raiz morfológica da palavra ou a forma de "dicionário da palavra" ou lema da palavra. Diferentemente do processo de *Stemming,* a lematização depende da identificação correta da POS da palavra e do significado da mesma, já que um algoritmo de *Stemming* opera em uma única palavra sem conhecimento do contexto e, portanto, não consegue fazer a discriminação de palavras que têm diferentes significados dependendo do contexto.

2.2.1.3.14 Reconhecimento de Entidades (REN)

Esta tarefa visa detectar entidades no texto, as quais podem ser pessoas, locais e instituições. Esta tarefa é interessante quando se busca identificar potenciais alvos de um sentimento específico.

2.2.1.3.15 Extração de Aspectos

Detecção de aspectos referentes a entidades em um texto. Por exemplo, um aparelho celular tem sua tela e botões para ligar e bloquear o aparelho, estes elementos são aspectos da entidade celular. Esta tarefa é adotada quando se aplica AS em nível de conceito.

2.2.1.3.16 Análise de Contexto

Nesta tarefa, almeja-se identificar um potencial contexto externo para um entendimento mais profundo do texto a ser analisado. Por exemplo, caso entidades nomeadas estejam presentes, busca-se entender qual a representação das mesmas para o significado daquele texto, tal como quem é a pessoa identificada ou com o quê a instituição detectada trabalha.

2.2.1.3.17 Análise Semântica e Sintática

Além de considerar o contexto semântico, esta tarefa busca identificar as relações de dependência entre diferentes termos dentro de um texto. Por exemplo, quem é o sujeito, o verbo e o objeto dentro de uma oração.

Assim, a escolha sobre quais tarefas serão aplicadas em uma dada base textual, depende da natureza da mesma. Por exemplo, quando se lida com dados formais, no caso de notícias, não se espera encontrar características como informalidades e erros de ortografia. Dessa forma, tarefas focadas em limpeza e redução dos dados podem ser aplicadas, tais como conversão para minúsculas e *Stemming* (PROLLOCHS; FEUERRIEGEL; NEUMANN, 2015).

Após pré-processar a coleção de documentos, o processo de MT segue para a transformação dos dados. A Subseção 2.2.1.4 a seguir apresentará um modelo de representação textual, o qual é o mais utilizado na literatura e que serve como embasamento para a representação utilizada neste trabalho.

2.2.1.4 Modelo para Transformação de Dados Textuais

Os modelos computacionais e algoritmos de TM em geral, necessitam trabalhar com dados textuais a partir de uma representação transformada dos mesmos. Nesse contexto, a literatura revela modelos de representação computacional de dados textuais. Dentre eles, tem-se o mais adotado, que é o *Bag-of-Words*. Este provê uma representação

numérica a partir de dados textuais, para modelos computacionais de processamento de linguagem natural.

No modelo de *Bag-of-Words* (ZHANG; LECUN, 2015), a representação do texto dá-se a partir de um vetor contendo pesos para cada termo ou palavra única encontrados em toda a coleção textual. Desta forma, este vetor terá o tamanho equivalente ao vocabulário do *corpus* disponível. Assim, cada célula do vetor contém o valor de peso para cada palavra dentro de toda a coleção textual.

Os pesos adotados dentro deste vetor são variados. O exemplo mais simples considera a ocorrência ou não do termo dentro de um determinado texto, sendo denominado Presença do Termo (do inglês, *Term Presence* - TP). Neste esquema, cada célula do vetor conterá o valor 1, caso o termo esteja presente, ou 0, caso o termo esteja ausente na respectiva instância analisada.

Um segundo exemplo de peso é denominado a frequência com que cada termo ocorre dentro de cada instância da coleção textual. Neste caso, cada célula do vetor irá conter o número de vezes que o termo se repete, dentro de um determinado texto. Se o mesmo está ausente, o valor é 0. Este esquema é denominado Frequência do Termo (do inglês, *Term Frequency* - TF) que é dado pela Equação 1.

$$tf_{i,j} = f_{i,j} \,, \qquad \text{Eq. (1)}$$

onde $tf_{i,j}$ representa o peso do termo *i* no documento *j* e $f_{i,j}$ representa a frequência do termo *i* no documento *j*.

Normalmente ocorre a normalização para valores no intervalo [0,1] para solucionar problemas associados ao tamanho do documento, evitando-se que um termo tenha uma frequência maior simplesmente porque o documento é grande. Para cada documento divide-se o valor da frequência de cada termo pela frequência mais alta do documento considerado (LOPES, 2004).

O esquema considerado mais sofisticado para essa atribuição de pesos é denominado de Frequência do Termo pela Frequência Inversa de Documentos (do inglês, *Term Frequency – Inverse Document Frequency* - TF-IDF). Nesta representação, o peso de cada termo considera a importância do mesmo em todos os documentos na coleção. O mesmo aumenta de acordo com a frequência do termo em um documento e com a sua raridade em toda a coleção. Dessa forma, o esquema TF-IDF tenta atribuir altos pesos

para palavras raras em toda a base, porém frequentes dentro de um documento. A forma de calcular este peso é ilustrada na Equação 2.

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right), \quad \text{Eq. (2)}$$

onde, $tf_{i,j}$ representa o total de ocorrências do termo i no documento j, df_i representa o total de documentos que possuem o termo i e N representa o número total de documentos.

Desta forma o peso $tfidf_{i,j}$ será:

- Alto quando o termo t ocorrer várias vezes em um número pequeno de documentos, apresentando assim um alto poder discriminante nesses documentos;
- Baixo quando o termo ocorrer poucas vezes em um documento ou ocorrer em vários documentos, apresentando assim um sinal menos acentuado de relevância estatística para Aprendizado de Máquina.

Após a transformação da coleção textual para a representação supracitada, é possível compor o grupo de vetores de entrada para os algoritmos de MT.

2.2.1.5 Tarefas de Mineração de Texto

Há diversas técnicas e tarefas de MT reportadas na literatura, tal como o processo de KDD, as tarefas de MT podem utilizar algoritmos de ML e algoritmos especializados em para a identificação e extração dos padrões presentes nos dados textuais. Os algoritmos de MT podem ser classificados como algoritmos de aprendizado supervisionado e não supervisionado.

O primeiro grupo de algoritmos visa encontrar um modelo computacional, gerado a partir de dados de treinamento rotulados (atributos de entrada), que possa ser utilizado para prever um rótulo ou valor para novas amostras de dados. Estes algoritmos são classificados pelo tipo dos rótulos dos dados: discreto, para tarefas de classificação; e contínuo, para tarefas de regressão (CARVALHO; *et al*., 2011).

Já os algoritmos de aprendizado não supervisionado visam explorar e descrever o conjunto de dados de entrada e não possuem saída. Geralmente são classificados em tarefas de: agrupamento, onde os dados são agrupados de acordo com a similaridade entre

os mesmos; sumarização, onde o objetivo é encontrar uma simples e compacta representação do conjunto de dados; e associação, que consiste em encontrar padrões frequentes entre os atributos de um conjunto de dados (CARVALHO; *et al*., 2011).

Dentre as principais tarefas de MT reportadas na literatura tem-se: Extração de Informação, Sumarização de Tópicos, Agrupamento de texto, Modelagem de tópicos, Classificação de texto, Tradução Automática e AS. O foco deste trabalho está sobre o processo de AS para PT_Br, então se faz necessário a apresentação dos principais conceitos de AS que serviram como embasamento para esta dissertação, os quais serão apresentados na Subseção 2.3.

2.2.1.6 Avaliação e Interpretação dos Resultados

Após a etapa de Mineração de Texto e avaliação dos resultados dos algoritmos, o próximo passo é interpretar os padrões minerados. Caso os resultados não estejam satisfatórios, pode-se retornar a qualquer um dos passos anteriores. Este passo pode também envolver a visualização dos padrões extraídos e modelos ou visualização dos dados.

Por fim, a última etapa atua sobre o conhecimento descoberto com a consolidação e validação de maneira que o conhecimento extraído possa ser validado por especialistas e posteriormente possa ser reportado às partes interessadas e incorporado aos processos ou sistemas de suporte à decisão. Esta etapa também inclui a verificação e resolução de potenciais conflitos com o conhecimento que se acreditava anteriormente.

2.3 ANÁLISE DE SENTIMENTO

Análise de Sentimento ou Mineração de Opinião consiste em classificar documentos de texto baseando-se no sentimento presente nos mesmos (PANG; LEE, 2008), ou seja, o sentimento que o autor deseja passar. Tal classificação é usualmente categorizada em classes positiva, negativa ou neutra. Para o surgimento dessa técnica, foram necessários avanços nas áreas de NLP e MT.

Esta subseção descreve um breve histórico e introduz os conceitos de AS, as estratégias utilizadas para a Mineração de Opinião e quais as métricas que podem ser usadas para a avaliação de desempenho de algoritmos classificadores de sentimento.

2.3.1 Breve Histórico

Análise de Sentimento tem sua base em NLP e MT e nos primórdios das pesquisas em fazer os computadores entenderem a linguagem humana (JONES, 1992). Durante os anos 70, linguistas trabalhavam fortemente no desenvolvimento da teoria gramatical. Além disso, havia uma tendência no uso da lógica para representação de conhecimento e raciocínio em Inteligência Artificial (IA).

Esses fatos impulsionaram linguistas a desenvolverem uma variada gama de tipos de gramática, que poderiam ser aproveitadas no contexto computacional quanto a estrutura de frases, como funcional, categórica e generalizada. Dado que existia também uma impulsão para o desenvolvimento da programação lógica, aliada ao desenvolvimento da teoria da gramática computacional (JONES, 1992).

A partir dos anos 80, a estatística começou a se destacar e ser mais considerada nos métodos em NLP. Essa fase também foi marcada pelo surgimento do dicionário léxico como uma base de conhecimento. Um dicionário léxico representa uma lista de palavras referentes a algum idioma, o qual também pode apresentar as categorias às quais essas palavras pertencem. Ao mesmo tempo, iniciativas na direção de coleção de dados e codificação dos mesmos, impulsionaram a tendência na utilização de *corpora* (plural de *corpus*) em aplicações de NLP.

Os avanços no desenvolvimento de equipamentos e computadores, no que concerne à capacidade de processamento e armazenamento, contribuíram para o surgimento de aplicações mais sofisticadas e robustas, presentes atualmente. Existe uma gama de *software* voltada especificamente para o processamento de linguagem natural, seja ela em forma de áudio, imagem ou texto. Como exemplo, tem-se o Natural Language Toolkit (NLTK[9]) e spacCy[10] para a linguagem de programação Python[11], além do ClearNLP[12] e CoreNLP[13] para a linguagem Java. Esse trabalho foca nos dados textuais, e ferramentas dessa área serão mencionadas e explicadas na Subseção 4.4.

[9] http://www.nltk.org/
[10] https://spacy.io/
[11] https://www.python.org/
[12] https://github.com/clir/clearnlp
[13] http://stanfordnlp.github.io/CoreNLP/

Nas últimas duas décadas, além dos fatos supracitados, o número de pesquisas em AS aumentaram consideravelmente (PIRYANI; MADHAVI; SINGH, 2017) devido também ao crescimento e disponibilização de bases de dados textuais, graças à expansão do uso da Internet e das plataformas da Web 2.0, que permitiram a geração massiva de UCG. Destaca-se o interesse comercial da indústria nos avanços das pesquisas e técnicas de AS, para obtenção de *insights* sobre os consumidores a fim de melhorar as estratégias de negócios (PANG; LEE, 2008).

2.3.2 Estado da Arte em Análise de Sentimento

Para a execução da técnica de AS, alguns passos são usualmente adotados na literatura. Além disso, existem variadas metodologias de implementação da Mineração de Opinião em si, bem como para se representar dados textuais. As aplicações desta técnica também são diversas, tanto para indústria quanto academia. Esta subseção irá focar na apresentação desses aspectos.

- Etapas Da Análise De Sentimento

As etapas para a aplicação da técnica de AS remetem aos passos padrões encontrados no *pipeline* de Mineração de Texto (FELDMAN; DAGAN, 1995)(REZENDE, 2005) apresentados na Subseção 2.2.1. Dentre estas, tem-se os estágios de coleta de dados, pré-processamento, transformação dos dados, processamento dos dados e avaliação de modelos.

No contexto de AS para dados de RSO, a aquisição dos mesmos é usualmente feita através de *Application Program Interfaces* (APIs) (ARUN; NAYAGAM, 2014). Por exemplo, plataformas como Twitter[14] e Facebook[15] possuem suas APIs, das quais é possível extrair, por exemplo, *tweets* ou comentários em *fan pages.* Em geral, estas APIs retornam a resposta em formato JSON, com o qual o trabalho de extração de dados é facilitado. Outra opção de coleta de dados é através de bases estáticas ou *off-line*. Existem sites especializados em hospedar tais coleções, como é o caso de *data.world*[16] e o *UCI Machine Learning Repository*[17].

[14] https://developer.twitter.com/en/docs.html
[15] https://developers.facebook.com/docs/graph-api/
[16] https://data.world/
[17] https://archive.ics.uci.edu/ml/index.php

A etapa de pré-processamento é aplicada para o melhor tratamento de dados, antes da transformação dos mesmos para modelos de classificação de sentimento. Esta engloba tarefas de pré-processamento padrão de MT descritas na Subseção 2.2.1.3 e outras tarefas específicas que atuam na limpeza e normalização de dados textuais, as quais lidam com a linguagem fora do padrão formal da língua que é utilizada em postagens em RSO e avaliações de produtos.

Existem diversas metodologias de implementação das tarefas de pré-processamento de dados textuais para AS. Uma delas é a Substituição de termos específicos por anotações ou identificadores únicos. Por exemplo, todas as ocorrências de palavras positivas podem ser substituídas por um identificador "positive" no texto (BALAHUR; PEREA-ORTEGA, 2015). Demais estratégias incluem a ação de Remoção (RM), que remete a simples deleção de um elemento. A Colocação (CO) trata da correta colocação ou inserção de um elemento no texto. A Computação de Polaridade (CP) é a ação de considerar diretamente o impacto de um elemento no sentimento final, seja intensificando-o ou computado-o. A Correção (CR) é notada para as tarefas de correção ortográfica. A Transformação (TR) remete a tarefas que tem seu modo particular de execução, e realizam alguma transformação nos dados, tal como tokenização e *stemming*.

A Tabela 2.1 ilustra exemplos de aplicações das metodologias previamente citadas.

TABELA 2.1 EXEMPLOS DE METODOLOGIAS PARA TAREFAS DE PRÉ-PROCESSAMENTO.

Metodologia	Entrada	Saída
Substituição	Joao eu amo muitooo esse filme!!! Vingadores é mto bon! :D	Joao eu positive muitooo esse filme!!! Vingadores é mto bon! :D
Remoção		Joao eu amo muito esse filme é mto bon
Colocação		Joao, eu amo muitooo esse filme!!! Vingadores é mto bon! :D
Computação de Polaridade		Polaridade = +1 (amo)
Correção		João eu amo muitooo esse filme!!! Vingadores é muito bom! :D
Transformação		Joao eu am muit ess film!!! Vingadores é mto bon! :D

No exemplo acima, a Substituição troca a palavra "amo" pela anotação "positive". A metodologia de Remoção atua deletando elementos da entrada, tais como repetições de letras em "muitooo", sinais de exclamação, e o *emoticon* ":D". Já a metodologia de Colocação, atua na correta inserção da vírgula após o nome "Joao". A Computação de Polaridade atribui pontos positivos de acordo com elementos encontrados, neste caso a

palavra positiva “amo”. A Correção insere o acento til em “Joao”, e corrige a palavra “bon” para “bom”. Por fim, a transformação neste exemplo é representada pela tarefa de *Stemming*, que atua na normalização das palavras “am”, “muit”, “ess” e “film”.

Assim, a literatura reporta diversas tarefas de pré-processamento de texto que podem ser aplicadas do contexto de dados de RSO. Neste âmbito, serão apresentados a seguir os conceitos de tarefas de pré-processamento utilizadas na proposta desta dissertação. Estas tarefas também podem ser adotadas no âmbito da MT em geral, porém são mais específicas para a aplicação da subtarefa de MT denominada Análise de Sentimento.

As mesmas estão organizadas em dois tipos: i) tarefas de identificação de conteúdo textual com sentimento neutro, ou que intensifiquem, reduzam ou invertam um sentimento; ii) tarefas que tratem elementos que capturam sinais diretamente positivos ou/e negativos. Além disso, os termos adotados para identificar cada tarefa nos capítulos seguintes, estarão presentes entre parênteses.

Tarefas de Pré-processamento Neutras

- Identificação de Amplificadores (Amplificadores): Esta tarefa de pré-processamento tem como objetivo identificar palavras que podem intensificar um dado sentimento, tais como advérbios de intensidade. Exemplos são: “muito”, “bastante” e “demais”;
- Identificação de Detratores (Detratores): Esta tarefa de pré-processamento almeja identificar palavras que têm o poder de diminuir a intensidade de um sentimento. Esta também inclui advérbios de intensidade, tais como “menos”, “pouco” e “apenas”;
- Identificação de letras maiúsculas (Maiúsculas): Esta tarefa de pré-processamento considera letras maiúsculas ou palavras escritas com letras maiúsculas como sinais de intensificação de sentimento. Por exemplo, “Eu AMO esse filme!” pode ser mais positiva que “Eu amo esse filme!”;
- Identificação de repetição de letras (Repetições de Letras): Esta tarefa de pré-processamento considera letras repetidas como sinais de intensificação e

sentimento. Por exemplo, “Eu amoooo esse livro!” pode ser mais positiva que “Eu amo esse livro!”;

- Remoção de repetições de letras (Remoção de Repetições de Letras): Esta tarefa de pré-processamento remove letras repetidas de palavras. Esta tarefa pode ser considerada também como uma forma de correção ortográfica (BALAHUR; PEREA-ORTEGA, 2015);
- Identificação de exclamação (Exclamação): Esta tarefa de pré-processamento considera um sinal de exclamação como um sinal intensificador de um dado sentimento;
- Repetição de exclamação (Repetições de Exclamação): Esta tarefa de pré-processamento considera sinais de exclamação repetidos como um sinal intensificador de um dado sentimento;
- Identificação de interrogação (Interrogação): Esta tarefa de pré-processamento considera um sinal de interrogação para anular sentimento, ou somente para anotação de sua ocorrência;
- Repetição de Interrogação (Repetições de Interrogação): Esta tarefa de pré-processamento considera sinais de interrogação repetidos como um sinal intensificador de um dado sentimento, ou como anulador de sentimento;
- Tratamento de negação (Negação): Esta tarefa de pré-processamento considera partículas de negação como inversores de polaridade dentro de um texto. Por exemplo, na frase “eu não gosto de sorvete”, esta tarefa de pré-processamento deve ser capaz de capturar o papel da partícula “não” e classificar este texto como negativo para o fato de o usuário não gostar de sorvete. Em geral, algoritmos de AS adotam listas com termos com poder de negação, a fim de capturar tal sinal dentro de um texto;
- Identificação de advérbio de dúvida (Dúvida): É um sinal também considerado como detrator de sentimento. Quando um advérbio de dúvida, tal como “talvez”, é encontrado no texto, este tem potencial de ser um sinal de redução da intensidade de um sentimento;

- Tratamento de *Hashtags* (*Hashtags*): *Hashtags* são elementos adotados para indexar palavras chave ou tópicos, uma função criada na rede social Twitter. Porém, outras plataformas como o Facebook e Instagram já passaram a adotar este elemento. Estas são tratadas ou por remoção, substituição por um identificador, substituição pelo seu conteúdo, ou tratamento e separação de conteúdo, quando compostas por mais de um termo;

- Tratamento de menções (Menções): Estas são referências a outros usuários no contexto de uma RSO. Em geral, são tratadas pela remoção ou substituição por um identificador que indique a presença de uma menção;

- Tratamento de URLs ou *Hyperlinks* (URL): Representam ligações, conexões ou *hyperlinks* para outra página ou endereço na Internet. Em geral, são tratadas pela remoção ou substituição por um identificador que indique a presença de uma URL.

Tarefas de Pré-processamento Positivas ou/e Negativas

- Tratamento de gírias (Gírias): O tratamento de gírias pode ser realizado de formas variadas, de acordo com a literatura. Por exemplo, as mesmas podem ser convertidas para o padrão culto da língua ou terem suas polaridades específicas anotadas, para etapas posteriores da AS. Neste sentido, suas ocorrências podem ser substituídas pela informação afetiva que carregam;

- Tratamento de xingamentos (Xingamentos): Esta tarefa de pré-processamento considera xingamentos como sinais de informação afetiva. Em geral, podem ser considerados como possuindo polaridade negativa;

- Tratamento de *Emoticons* (*Emoticons*): Emoticons são representações tipográficas de expressões faciais, usadas para representar emoção em um meio de informação unicamente textual. Por exemplo, podem representar uma face feliz ou triste. Estes podem ser considerados como ruídos, e neste caso serem removidos do texto. Por outro lado, os mesmos podem ter seus sentimentos substituídos no texto, ou anotados para computação final da polaridade textual;

- Tratamento de *Emojis* (*Emojis*): *Emojis* ilustram emoções de forma lúdica dentro de um texto. Estes são realmente imagens das expressões, que são ilustradas somente por símbolos no caso de *emoticons*. O tratamento para estes sinais segue o mesmo padrão de *emoticon*, ou seja, podem ser removidos, substituídos por identificadores de sentimento, ou serem utilizados para computação da polaridade final;

- Tratamento de risadas (Risadas): Esta tarefa visa identificar padrões de risada ou gargalhada no texto, tal como “kkkk”, “hehehe”, “haha”. Neste caso, o tratamento pode ser feito pela remoção ou substituição desses elementos por um identificador, que indique a presença de risadas;

- Tratamento de cumprimentos (Cumprimentos): Esta tarefa busca detectar cumprimentos no texto, tal como “bom dia”, “boa tarde” e “boa noite”. Os mesmos podem ser tratados por remoção ou substituição desses elementos por um identificador, que indique a presença de cumprimentos;

- Tratamento de palavras afetivas (Palavras Afetivas): Esta tarefa visa identificar palavras com conotação positiva ou negativa no texto. Estas são geralmente tratadas pela substituição por uma anotação, que indique qual o sentimento das mesmas.

Após a aplicação das tarefas de pré-processamento, o *pipeline* de AS segue para estágio de transformação dos dados de um domínio textual tal como no processo de MT, para um domínio onde estes podem ser representados numericamente, para modelos computacionais de processamento de texto. O modelo utilizado neste trabalho é o modelo de *Bag-of-Words*, o qual foi explanado anteriormente.

A etapa de processamento dos dados em AS faz referência, ao processo de classificação de sentimento propriamente dito. Para este, existem metodologias de referência no estado da arte. As mesmas serão ilustradas na Subseção 2.3.3 a seguir.

Após a execução de um *framework* de AS, os resultados podem ser analisados. Para isto, existem métricas de avaliação comumente adotadas nesta literatura, as quais também são originadas da avaliação de modelos de aprendizado de máquina em geral. Estas serão detalhadas na Subseção 2.3.5.

2.3.3 Metodologias para Análise de Sentimento

As metodologias de AS em geral, variam de acordo com a estratégia e modelo escolhido para aplicação desta técnica. Em (PORIA, *et al*., 2014), tem-se a definição de quatro estratégias para AS, as quais são comumente encontradas em diversos trabalhos na literatura: *Keyword spotting,* afinidade léxica, estatística e Aprendizado de Máquina e Nível de conceito.

Algumas abordagens apresentam também estrutura híbrida entre as propostas supracitadas, a fim de suprir um problema de outras estratégias. Em (RIBEIRO, *et al.*, 2010), encontra-se um *survey* do estado da prática com vinte e dois dos mais utilizados métodos de AS. Os autores realizaram uma avaliação de desempenho dos algoritmos com dados de RSO, *reviews* e comentários de blogs. Além disso, essas são geralmente as fontes de dados mais utilizadas por propostas da literatura.

1. A primeira metodologia, e a mais comum para AS, é o *keyword spotting* (PORIA, *et al*., 2014), que tem como base o uso de palavras afetivas e com polaridade positiva ou negativa, para detectar o sentimento do texto. Esse método pode ser considerado o mais ingênuo, pois considera unicamente a presença de termos com alguma afetividade para sua decisão.

A abordagem *keyword spotting* apresenta como vantagem a sua fácil acessibilidade, pois não é um método complexo para se implementar, ou seja, sem muito custo. A desvantagem dessa estratégia está em sua superficialidade, pois tem como base características explícitas no texto, como as palavras com alguma carga sentimental, tais como “ótimo”, “ruim”, e “legal”. Caso um texto não apresente alguma palavra deste tipo, a classificação pode ser prejudicada. Além disso, eventos como negação podem não ser tratados, o que pode gerar erros nos resultados, dado que isso é um fator muitas vezes determinante em textos com opiniões (PORIA, *et al*., 2014).

2. A segunda estratégia mencionada por (PORIA, *et al*., 2014) faz referência ao método de afinidade léxica. Dessa vez, a análise conta com um fator de maior robustez, que não é somente a presença ou ausência de termos indicadores de sentimento. Essa maior sofisticação vem do aprendizado através de corpora para compor probabilidades atribuídas a termos, quanto à presença destes em textos com polaridade positiva ou negativa.

Entretanto, o método de afinidade léxica apresenta como desvantagem a falta de robustez quanto à negação. Por exemplo, a palavra “acidente”, no geral, aparece em um contexto negativo. Todavia, há a possibilidade de frases como “conheci minha namorada por acidente”, a qual não apresenta uma conotação negativa de fato. Porém, devido às probabilidades prévias atribuídas às palavras, as mesmas podem ter tendência maior a aparecerem em um determinado tipo de contexto, o que pode prejudicar os resultados da classificação.

3. A terceira metodologia é a estatística, a qual tem relação com técnicas e algoritmos de ML, tais como Redes Neurais Artificiais (HAGAN, 1996), Support Vector Machine (CORTES; VAPNIK, 1995) e inferência bayesiana (BOX; TIAO, 2011). Nesse caso, usualmente trabalha-se com um corpus de treino, contendo dados que foram avaliados previamente, de forma manual. Em seguida, esses dados são fornecidos para modelos de aprendizado de máquina, que então irão se especializar na classificação de acordo com os dados de treino fornecidos.

Essa metodologia geralmente apresenta uma boa acurácia quando se utiliza uma grande quantidade de dados de treino (PORIA, *et al*., 2014). Além disso, essa estratégia pode ser boa para a classificação em nível de página ou parágrafo.

Como desvantagens este terceiro método tem: 1) não apresenta robustez quanto à análise semântica, pois não leva em consideração o papel de cada termo no texto, mas apenas as suas estatísticas quanto à presença ou não em documentos; 2) com a exceção de palavras com alguma característica sentimental, outras palavras não apresentam um grande poder de predição de sentimento nesse modelo (PORIA, *et al*., 2014). Por fim, esse método pode apresentar performance baixa ao classificar sentenças curtas.

Diversos classificadores são utilizados na literatura em AS, porém o algoritmo *Support Vector Machine* (SUN; LUO; CHEN, 2017) destaca-se pelos resultados superiores em relação a outros algoritmos em tarefas de AS (SOUZA, ELLEN, *et al*, 2016), conforme é reportado no estado da arte. Assim, este trabalho faz uso do algoritmo SVM como classificador.

A Máquina de Vetores de Suporte ou *Support Vector Machine* é um algoritmo de classificação embasado na Teoria de Aprendizado Estatístico (TAE) (VAPNIK,1998)(VAPNIK, 1995)(VAPNIK; CHERVONENKIS, 2015 *apud* VAPNIK;

CHERVONENKIS, 1971) e que destacam-se na literatura pela boa capacidade de generalização (CARVALHO; *et al*, 2011).

A primeira formulação proposta de SVM é capaz de lidar com problemas linearmente separáveis (BOSER; *et al*., 1992), com a adoção de fronteiras de decisão lineares ou *kernel* linear. Essa formulação foi estendida para conjuntos de dados mais gerais (CORTES; VAPNIK, 1995), a partir de diferentes *kernels* para tomada de decisão quanto a predição de classes. Assim, o SVM possui quatro tipos principais de *kernels*, a saber: linear, radial, polinomial e sigmoid (MEYER; WIEN, 2001).

4. A proposta de análise em nível de conceito pode ser considerada a mais sofisticada dos métodos para AS, pois nesta não são levadas em consideração somente palavras chave, termos isolados ou análises estatísticas, mas o significado dos termos (PORIA, *et al*., 2014). Dessa forma, a análise semântica dos termos é realizada com utilização de ontologias web ou redes semânticas de palavras, geralmente encontradas em forma de ferramentas de software prontas (PORIA, *et al*., 2014), tais como o Onto.PT (OLIVEIRA; GOMES, 2014). Essas plataformas realizam a agregação de informação conceitual e afetiva.

Assim, a base dessa metodologia não está em características explícitas do texto, como quais termos ocorrem, mas nas entrelinhas e informações implícitas associadas a cada termo de um documento, como suas formas de associação com outros termos. Como exemplo, podem aparecer palavras em um texto que não tem uma conotação positiva ou negativa. Entretanto, as mesmas podem ter relação com outros termos que tenham uma polaridade determinada. Então, utilizando-se essa informação, a acurácia de métodos de AS poderia ser melhorada.

A AS em nível de conceito tenta ainda englobar estratégias referentes ao próprio funcionamento do cérebro humano, como por exemplo, o uso do senso comum para tomar decisões. No exemplo da frase “sala pequena”, o termo “pequena” pode ter conotação negativa, para o caso de uma *review* de hotel. Entretanto, na frase “fila pequena”, a mesma palavra tem conotação positiva, dado que uma fila com poucas pessoas geralmente é preferencial em qualquer cenário. A metodologia em nível de conceito visa obter a classificação correta nesses casos também, por meio do senso comum e ontologias (PORIA, *et al*., 2014).

As metodologias de AS podem ainda variar quanto ao escopo da classificação do sentimento. Em (LIU, 2012), é possível observar a definição do escopo de AS em três níveis principais: nível de documento, nível de sentença e nível de aspectos.

A classificação em nível de documento é a forma mais simples de AS e assume que o documento contém uma opinião sobre um objeto principal expresso pelo autor do documento (FELDMAN, 2013). Este nível de AS considera todo o texto como contendo uma polaridade definida. Por exemplo, este nível pode considerar um *tweet* com a polaridade positiva, representada por todo o seu conteúdo textual ou um comentario do Facebook como tal.

Na classificação ao nível de sentença, como o nome sugere, leva-se em conta cada sentença dentro de um documento para fazer a classificação. Assim, essa metodologia assume que sentenças diferentes de um texto inteiro podem apresentar polaridades diferentes. Este nível de AS considera que um documento pode ter várias opiniões sobre as mesmas entidades. (FELDMAN, 2013)

Por fim, a classificação em nível de aspectos é a de maior complexidade, pois se concentra no reconhecimento de todas as expressões de sentimento dentro de um determinado documento e os aspectos as quais se referem. Um aspecto é um atributo de uma determinada entidade que pode estar explícito ou implícito no texto. Normalmente, aplica-se este nível de classificação quando se deseja avaliar o sentimento para cada aspecto de uma entidade. Por exemplo, pode-se utilizar a classificação em nível de aspectos em comentários e avaliações sobre *smartphones*, vários atributos ou aspectos de categorias de produtos são avaliados como processador, câmera, memória RAM, armazenamento, conectividade, duração da bateria etc. (FELDMAN, 2013).

2.3.4 Aplicações de Análise de Sentimento

As aplicações da técnica de AS podem ser encontradas em uma diversidade de domínios como, por exemplo, precificação de produtos, inteligência competitiva, predição de eleições, detecção de risco em sistemas bancários, e no setor de saúde. Todas estas são aplicações que tem atraído à atenção da indústria e academia.

Na área de precificação de produtos, o trabalho de (ARCHAK; GHOSE; IPEIROTIS, 2011) analisou revisões de produtos do site *Amazon.com* através de AS, para

capturar as características preferidas de consumidores em produtos eletrônicos. Os autores almejavam uma predição de vendas futuras, de acordo com os comentários de tais usuários. Assim, o preço de produtos comentados poderia ser configurado de acordo com estas predições.

No setor de inteligência competitiva, a proposta de (XU, KAIQUAN, *et al.,*2011) ilustrou um exemplo onde comentários de usuários sobre diferentes marcas e produtos, eram monitorados. A ideia era aplicar Mineração de Opinião, a fim de verificar em quais aspectos consumidores estavam reclamando ou provendo elogios sobre competidores. Com esse panorama, empresas poderiam ter o poder de melhorar suas ofertas e potencialmente alterar estratégias de *marketing*, para superar concorrentes.

O período de eleições acaba por se configurar como uma época com oportunidades de negócios, principalmente para candidatos ao pleito. A área de AS também já foi aplicada neste setor. Por exemplo, o estudo de (CERON, ANDREA, *et al*.,2014) aplicou esta técnica, no contexto político, no cenário de mídias sociais em dois países. O primeiro, foi a Itália, onde o autor tentou identificar a popularidade de líderes políticos. O segundo, foi a França, no qual o estudo tentou capturar a intenção de votos a partir dos sentimentos de eleitores em plataformas de mídia social. Os resultados ilustraram, por exemplo, que existe uma correlação entre comentários em mídias sociais e pesquisas de intenção de votos.

Outra aplicação observada de AS foi no setor bancário, através do estudo de (NOPP; HANBURY, 2015). Neste, foram analisadas mais de 500 cartas e relatórios anuais de diferentes bancos europeus, a fim de se supervisionar por meio do sentimento identificado em tais documentos, a atitude de bancos quanto a possibilidades de riscos para seus negócios e para a economia como um todo.

No setor de saúde, o trabalho de (RODRIGUES; RAMON GOUVEIA, *et al*., 2016), visou o monitoramento do estado emocional de pacientes com câncer, em comunidades de RSO sobre esta temática. Para tal objetivo, os autores empregaram AS a fim de identificar os sentimentos expressados por pacientes em seus comentários. Dessa forma, seria possível otimizar o atendimento a estes, com ações imediatas quando uma crise fosse identificada em alguns dos pacientes, relacionada a seu estado e atitude quanto a este problema.

2.3.5 Avaliação de Métodos de Análise de Sentimento

A avaliação de algoritmos de AS se configura como a última etapa dos estágios de execução desta técnica. Sua importância está na verificação do desempenho de modelos adotados nesta temática, a fim de relevar o nível de eficiência dos mesmos.

As métricas comumente utilizadas na literatura para demonstração de performance de métodos de Mineração de Opinião e que foram utilizadas nesse trabalho, são apresentadas a seguir (RIBEIRO, *et al*., 2010).

- Acurácia: uma das formas mais intuitivas de se acessar a performance. É definida como a razão entre o total de predições corretas e o total de predições observadas;
    - $$\text{Acurácia} = \frac{\text{Total de Predições Corretas}}{\text{Total de Predições Observadas}} \quad (2.1)$$
- Precisão: a precisão é a acurácia, porém com ênfase em uma das classes de classificação. É a razão entre o número de predições corretas para uma classe pelo número total de predições observadas para aquela classe;
    - $$\text{Precisão (X)} = \frac{\text{Total de Predições Corretas da Classe X}}{\text{Total de Predições Observadas da Classe X}} \quad (2.2)$$
- Revocação: esta métrica também é conhecida como sensitividade, verdadeira taxa de positivos, a qual dá a proporção total de verdadeiros positivos. A revocação ou *recall* é definida como o número de elementos corretamente classificados como uma classe, pelo total de elementos que são verdadeiramente daquela classe;
    - $$\text{Revocação (X)} = \frac{\text{Total de Predições Corretas da Classe X}}{\text{Total de Elementos da Classe da X}} \quad (2.3)$$
- F1: esta métrica representa a média harmônica entre precisão e revocação, definida por:
    - $$\text{F1 (X)} = \frac{2\text{Precisão(X)} \,.\, \text{Revocação(X)}}{\text{Precisão(X)} + \text{Revocação(X)}} \quad (2.4)$$

Onde:

i. X é uma das classes.

- Macro F1: uma visão geral sobre a F1 para cada classe presente nos experimentos. Assim, a Macro F1 se caracteriza como a média das F1 adquiridas, sobre todas as classes. Esta métrica é importante, pois, pode prover uma análise justa sobre o desempenho de modelos de AS na predição de variadas classes, mesmo quando uma base de dados não é balanceada.

  - $\text{Macro F1} = \frac{\text{F1(X)+F1(Y)+}\cdots\text{+F1(N)}}{\text{Total de Classes}}$ (2.5)

Onde:

i. X, Y a N são classes presentes na base de dados.

Demais avaliações realizadas nesta dissertação incluem análises quanto ao desempenho computacional geral e individual para cada tarefa de pré-processamento, além das tarefas empregadas para a maior precisão em cada base de dados.

2.4 A LINGUAGEM EM REDES SOCIAIS

Rede social é um conceito anterior ao surgimento da *Internet*. Uma rede social faz referência a, por exemplo, um grupo de pessoas na sociedade que estão ligadas por interações sociais (RECUERO, 2009). Entretanto, com a disseminação da *Web* 2.0, que impulsionou o surgimento de plataformas interativas, também surgiram as mídias e Redes Sociais Online.

Formalmente, uma rede social tem sua definição baseada em dois elementos principais: atores e suas conexões. Os atores são considerados os membros, instituições ou grupos envolvidos, enquanto que as conexões são as interações ou laços sociais entre os mesmos, como afirmam (WASSERMAN; FAUST, 1994) e (DEGENNE; FORSE, 1999).

Para esta dissertação, duas plataformas de rede social na Internet foram de maior importância. A primeira foi o Facebook, o qual tem sua origem no ano de 2004, criado por Mark Zuckerberg, com seus colegas de quarto e antigo parceiro Eduardo Saverin (PHILLIPS, 2007). A segunda e não menos importante, foi o Twitter. Essa plataforma de *microblog* foi criada em Março de 2006 por Jack Dorsey, Evan Williams, Biz Stone e

Noah Glass. Porém, o site foi lançado somente em Julho do mesmo ano (TWITTER MILESTONES, 2016).

Quanto ao Facebook, segundo (RECUERO, 2014a), os seus atores podem ser considerados os próprios usuários e seus perfis pessoais, ou as *fanpages*, que podem ser representações institucionais, de ideologias, celebridades ou usuários comuns. Em 2016, esta RSO reportou que contava com mais de 1.65 bilhões de usuários ativos mensalmente e 1.09 bilhões ativos diariamente, em diversos países (FACEBOOK COMPANY INFO, 2016). O volume de dados gerados também é grande nessa rede, sendo 300 milhões de fotos postadas por dia, 293 mil *statuses* atualizados e 510 mil novos comentários postados a cada 60 segundos (ZEPHORIA, 2018).

Enquanto isso, o Twitter foi fundado como um projeto da empresa Odeo[18]. Ele é considerado uma plataforma de *microblogging* (JAVA, *et al*., 2007), por sua natureza de postagens, as quais são principalmente textos de até 140 caracteres, denominadas *tweets*. Entretanto, essa plataforma também permite postagens de fotos e vídeos por seus usuários.

As últimas estatísticas do Twitter em 2016 indicavam que, o mesmo conta com mais de 1.3 bilhões de usuários registrados e 100 milhões de usuários ativos diariamente (SMITH, 2016). A exemplo do Facebook, seu fluxo de dados também é intenso. Segundo (INTERNET LIVE STATS, 2018), aproximadamente 6000 *tweets* são enviados a cada segundo, o que pode corresponder a 500 milhões de postagens por dia.

Quanto a comunicação nessas plataformas, a mesma é geralmente caracterizada pela velocidade e grande alcance das informações, além da informalidade presente por parte dos seus usuários. Além disso, é comum observar a presença da linguagem universal e globalizada, marcada por neologismo de sentido e de forma que muitos usuários utilizam nas Redes Sociais Online em que estão inseridos (GALLI, 2002).

Os conteúdos gerados por usuários em RSO, em geral, tem características informais, tais como: abreviações; gírias; termos dependentes de contexto; e marcados por indiferença quanto à gramática e ortografia (NAGARAJAN; GAMON, 2011).

[18] https://www.crunchbase.com/organization/odeo#/entity

Além disso, a forma com que os usuários falam na Internet é diferente da linguagem utilizada no mundo real, e ainda possui um nome: o "Internetês" (KOMESU; TENANI, 2009). Esse termo se refere ao português que é escrito na Internet.

Como já foi mencionado anteriormente, o processo de globalização aproxima pessoas e torna o fluxo de informações mais veloz, além de gerar uma tendência por uma linguagem universal, que atenda essas necessidades e acompanhe esse processo (GALLI, 2002).

O Internetês é reflexo dessa necessidade, e ele pode ser considerado como uma parte da metamorfose natural da língua (apud MARCONATO, 2006) para ser uma forma de se comunicar mais rápida, se economizando "tecladas" (CRYSTAL, 2005; CASTILHO apud MARCONATTO, 2006).

Essa forma de se comunicar se desvia da norma culta padrão da língua portuguesa, representada por uma simplificação da escrita e redução da ortografia (KOMESU; TENANI, 2009). Muitos consideram como sendo uma linguagem escrita onde se tem grande interferência da fala, ou como uma escrita fonetizada. (KOMESU, 2007). Para ilustrar isso, (KOMESU; TENANI, 2009) provê quatro exemplos claros dessa afirmação.

- "HMM": pode representar na conversação uma "correção auto-iniciada feita no mesmo turno" (MARCUSCHI, 1986), que se caracteriza como uma reelaboração do dizer, que também visa garantir o turno durante uma conversa.
- "AINN": também pode representar uma maneira de se corrigir e reelaborar o dizer. Porém, diferentemente do exemplo anterior, é possível observar a presença de "ai", e reconhecê-lo como uma palavra em português. Quanto ao "NN", esse representa o som nasal.
- "HEHE": atua como representação dos sons produzidos por uma risada, que também é visto como uma maneira de realizar interação com leitores/interlocutores (KOMESU, 2002).
- "GOXTO": este é um exemplo do caráter de regionalismo no Internetês, no qual temos a troca de "S" por "X". Nessa situação, o emprego formal seria o do "S", que seria mais aproximado ao som do sotaque paulista. A troca do "S" por "X" na palavra mencionada, é uma tentativa do autor do texto de se

aproximar da reprodução do sotaque carioca, no qual o som do “S” tem aproximação com o som da letra “X”.

Portanto, a partir dessa análise é possível inferir que o cenário de mídias sociais e RSO é desafiador para técnicas de AS, dadas as infinitas possibilidades de variação que adeptos do Internetês podem utilizar em sua escrita, o que não é possível de ser completamente coberto por modelos computacionais de AS. Além disso, a cada dia podem surgir novas expressões, dependentes de contexto e eventos que afetam a sociedade, o que também pode não estar presente na arquitetura dos algoritmos de AS.

2.5 CONSIDERAÇÕES FINAIS

Neste capítulo apresentaram-se brevemente alguns temas relevantes para esta dissertação, iniciando pela contextualização da área de Análise de Sentimento, com foco em seu histórico e desenvolvimento, incluindo as áreas de Mineração de Texto e NLP.

Posteriormente, foram apresentados alguns conceitos dos processos de Mineração de Opinião. Estes foram relacionados ao *pipeline* geral de execução, além de detalhes referentes a cada um dos estágios nesse processo, principalmente nas metodologias adotadas e transformações em dados textuais.

Estas informações, bem como os exemplos de aplicação em diferentes áreas, também contribuem para fornecer um panorama geral sobre a mineração de dados textuais, e sua importância no contexto de mídias sociais. Por sua vez, este é o cenário alvo principal de atuação da arquitetura implementada neste estudo, o SentiBR.

Além disso, foi apresentado o conceito de mídias e Redes Sociais Online, as quais se constituem como os cenários centrais de aplicação neste estudo.

Destacaram-se os desafios encontrados neste contexto para métodos computacionais de análise de conteúdo textual. Foi ressaltado que tais modelos enfrentam barreiras, devido às informalidades, regionalismos e não uso da norma culta em plataformas nesta cena. Todos estes conceitos servem como base para os próximos passos apresentados nesta dissertação, os quais passam a enfatizar como enfrentar tais desafios, a fim de se executar aplicações de Mineração de Texto em tal conteúdo, principalmente a AS.

Assim, o próximo capítulo apresenta trabalhos relacionados referentes à abordagem de AS em geral, mas com um foco na etapa de pré-processamento, tema investigado com maior ênfase neste estudo. Além disso, serão fornecidos detalhes quanto ao contexto global e brasileiro, no ensejo de propostas que tentam lidar com os desafios de mídias sociais na aplicação de Mineração de Opinião.

3 TRABALHOS CORRELATOS

3.1 CONSIDERAÇÕES INICIAIS

A pesquisa por trabalhos relacionados à Análise de Sentimento deu-se pela busca na literatura por trabalhos que apresentassem uma síntese da temática, não restringidos a uma língua específica. Em seguida, a pesquisa bibliográfica foi focada em recentes artigos científicos, principais periódicos, além de teses, dissertações na área de Mineração de Texto e processamento de linguagem natural, com ênfase em pré-processamento para AS para o Português do Brasil. Entretanto, trabalhos mais antigos também estão incluídos, devido à relevância para a literatura estudada.

Dessa forma, este capítulo apresenta uma visão geral dos estudos relacionados ao tema de AS para mídias sociais, explicitando os trabalhos julgados mais relevantes, primeiramente no cenário internacional. Posteriormente, uma ênfase é dada para trabalhos desenvolvidos para o PT_Br. A etapa de pré-processamento sempre está destacada, sendo o principal tema investigado nesta dissertação.

O *framework* desenvolvido neste estudo estará identificado em tabelas e resultados neste capítulo e nos seguintes, com o seu nome de SentiBR.

3.2 PRÉ-PROCESSAMENTO EM ANÁLISE DE SENTIMENTO

A etapa de pré-processamento tem como objetivo tratar, transformar e selecionar as melhores características de uma base textual de dados, para uma etapa posterior de classificação de sentimento naquela base. Tais tarefas envolvem etapas introduzidas na literatura de MT em geral, como é o caso de tokenização, remoção de *stopwords* e remoção de pontuações (MOH, MELODY, *et al*., 2015), (FELDMAN; SANGER, 2007) e (KANNAN; GURUSAMY, 2014).

No caso de propostas de AS voltadas especificamente para dados de redes sociais, normalmente as mesmas apresentam regras e heurísticas voltadas ao tratamento particular de desafios encontrados nestes dados. Como apresentado previamente na Subseção 2.4, a linguagem em tal cenário tem natureza informal e por muitas vezes, uma gramática fora do padrão culto da língua (BAHRAINIAN; DENGEL, 2013).

Durante o desenvolvimento desta dissertação, revisões da literatura foram realizadas a fim de se detectar quais tarefas de pré-processamento tem sido encontradas para AS, tanto no cenário nacional quanto global. Nestas buscas, enfatizou-se a exploração de trabalhos voltados a dados de mídias sociais.

Dentre estas tarefas que lidam com características informais com potencial indicação de um sentimento presente em um texto de RSO, tem-se: deteção de intensificadores que enfatizam sentimento (i.e. advérbios de grau, letras maiúsculas, repetições de pontuação), tratamento de *hashtags* e *emoticons* com polaridades específicas, tais como risos ou rostos tristes (HUTTO; GILBERT, 2014), (LEVALLOIS, 2013), (AVANÇO; BRUM; NUNES, 2016), (RODRIGUES, RAMON GOUVEIA, *et al*., 2016) e (SOLAKIDIS; VAVLIAKIS; MITKAS, 2014).

3.3 ANÁLISE DE SENTIMENTO PARA MÍDIAS SOCIAIS

No contexto da era do *Big Data* em plataformas de RSO, diversas técnicas têm sido apresentadas na literatura, com foco na mineração de tais dados para otimização de estratégias de negócios para empresas e instituições (FELDMAN, 2013). Dentre estes métodos, variadas propostas com foco em AS para dados de mídias sociais têm sido observadas. Prova disso é a estatística revelada por (PIRYANI; MADHAVI; SINGH, 2017), a qual ilustra que o Twitter é a segunda fonte de dados mais utilizada em pesquisas em AS, desde o ano de 2000 até 2015.

A ênfase no cenário de mídias sociais tem sido observada em trabalhos para diversas línguas, o que reforça a importância da Mineração de Opinião no contexto global para empresas, profissionais de marketing e pesquisadores (GIATSOGLOU, *et al*., 2017). Todavia, o trabalho de (PIRYANI; MADHAVI; SINGH, 2017) destaca que a maioria das propostas têm sido implementadas para a língua inglesa.

Entretanto, o principal desafio enfrentado por propostas de AS para este cenário é o próprio estilo de linguagem adotado (FERSINI; MESSINA; POZZI, 2016). Como destacado no Capítulo 2, a linguagem utilizada na Internet por vezes é caracterizada como um novo código de comunicação (KOMESU; TENANI, 2009). De fato, o conteúdo gerado por usuários na Web 2.0 e RSO apresenta desafios para técnicas de NLP em geral, incluindo AS (CLARK; ARAKI, 2011).

Dentre estes desafios, é possível mencionar erros de ortografia, pontuação arbitrária, letras maiúsculas, abreviações e gírias (PETZ, *et al.*, 2013). Todos estes aspectos influenciam o desempenho de classificadores de sentimento e impõem barreiras no avanço da área para RSO (PETZ, *et al.*, 2014).

3.3.1 Contexto Global

Em (SINGH; KUMARI, 2016), os autores focaram em experimentos com diferentes conjuntos de passos de pré-processamento, incluindo tratamento de gírias, correção ortográfica e *Stemming*. A base de dados adotada era composta de mensagens do Twitter com *emoticons*. Como conclusão, os autores observaram que um conjunto apropriado de tarefas para tratar conteúdo gerado por usuários em RSO, poderia melhorar o desempenho de AS neste contexto.

O trabalho de (MOSQUERA; GUTIÉRREZ; MOREDA, 2017) avaliou técnicas de normalização baseadas em um dicionário léxico, para AS em português e espanhol, para dados advindos de diferentes plataformas da Web 2.0. Eles analisaram que o processo de normalização é efetivo para textos informais. No caso de *corpus* formais, tais tarefas podem trabalhar bem em alguns casos. Entretanto, os mesmos adotaram apenas 6 tarefas em seu *pipeline* completo de pré-processamento, o que não são suficientes para lidar com todos os desafios em UGC.

No caso da língua inglesa, um tradicional e bem estabelecido exemplo de algoritmo na literatura é o *SentiStrength* (THELWALL, *et al.*, 2010), construído para textos curtos informais de mídias sociais. Este método implementa o tratamento de erros ortográficos, negações, *emoticons*, intensificadores e detratores de sentimento. No caso destes últimos, os autores consideram advérbios de grau e repetições de letras ou pontuação. Posteriormente, (HUTTO; GILBERT, 2014) apresentaram outra proposta baseada em heurísticas e dicionário, denominada Vader, que inclui passos como letras maiúsculas para intensificação de sentimento, tratamento de gírias e abreviações. Apesar de ambos os trabalhos implementarem acima de 10 tarefas de pré-processamento, os mesmos não capturam a intensificação de sentimento por repetições de pontuações.

Ainda no contexto da língua inglesa, (KHAN; BASHIR; QAMAR, 2014) propuseram um *framework* com estratégia de classificação híbrida, considerando três

dicionários de sentimento, além de tratamento de gírias e verificação ortográfica. Recentemente, o trabalho de (FERSINI; MESSINA; POZZI, 2016) objetivou analisar os sinais mais expressivos de sentimentos em mídias sociais, com o potencial de melhorar o desempenho de AS neste cenário. Dentre tais sinais, os autores investigaram *emoticon*s, expressões enfáticas, de onomatopeia e repetições de letras. Os autores realizaram experimentos com diversos modelos de aprendizado de máquina e também concluíram que tais sinais podem contribuir para resultados promissores. Todavia, os mesmos aplicaram um único formato de intensificação em seu *pipeline*, além de não detectarem negações.

A estratégia multilíngue de AS também tem sido apresentada. Para ilustrar, o trabalho de (TELLEZ; ERIC, *et al*.,2017b) propôs um *framework* com ênfase em tratar desafios de mídias sociais que sejam independentes ou dependentes de uma linguagem específica. Por exemplo, tarefas independentes incluíram o tratamento de *emoticons*, números, URLs e menções, acentos, e símbolos repetidos. As tarefas dependentes de linguagem foram *stemming*, negação, e remoção de *stopwords*. Os autores realizaram uma avaliação com n-gramas baseados em palavras e caracteres, para cinco linguagens diferentes. Os mesmos alcançaram os melhores resultados em quatro desafios internacionais. Entretanto, estes autores não levaram em consideração gírias e regionalismos locais, bem como intensificação por sinais de pontuação, os quais são elementos frequentemente detectados em RSO.

Em (BALAHUR; PEREA-ORTEGA, 2015), um trabalho inédito foi apresentado com foco em ferramentas e recursos de pré-processamento em AS, considerando particularidades do inglês e espanhol. A base de dados deste estudo era composta de *tweets*, com a qual os autores atingiram desempenho considerável para a tarefa. Os mesmos aplicaram tarefas como conversão para minúsculas, tratamento de gírias, normalização de letras repetidas, tratamento de palavras afetivas, negação, intensificadores de sentimento e tratamento de menções e *hashtags*. Porém este autor também não lidou com Gírias, Xingamentos, Correção Ortográfica e Exclamação.

Focando exclusivamente em linguagens de origem romântica (PEETERS, 2006), o que inclui espanhol, italiano e francês, algumas propostas também estão presentes na literatura. Em (TELLEZ; ERIC, *et al*.,2017a), os autores conduziram uma revisão com técnicas populares para se lidar com conteúdo informal e curto de mídias sociais. Os mesmos incluíram tarefas de correção ortográfica, tratamento de negação, normalização

de repetições, *emoticons*, *stemming* e lematização. Com um classificador SVM, eles atingiram melhoras significativas para duas bases de dados em espanhol.

Novamente no cenário espanhol, (VILARES; THELWALL; ALONSO, 2015) realizaram uma adaptação do algoritmo *SentiStrength* para a língua espanhola, incluindo um pré-processamento focado em tratar conteúdo gerado por usuários. Os mesmos atingiram desempenho significativo para uma base de *tweets* sobre o tema de política. Entretanto, não mencionam com clareza o tratamento de elementos de RSO, como *hashtags*.

No contexto italiano, (MUSTO; SEMERARO; LOPS; GEMMIS, 2015) realizaram uma análise semântica de dados de tempo real do Facebook e Twitter, incluindo a tarefa de AS. Incluído em seu *pipeline* de pré-processamento, estavam as tarefas de remoção de *stopwords*, *stemming¸ Part-of-Speech* e expansão de abreviações comumente encontradas em RSO italianas. Como principal objetivo, eles tentaram detectar opiniões de habitantes de uma cidade italiana.

Quanto ao português de Portugal, propostas também têm sido observadas. O estudo de (SAIAS; JOSE, *et al.*, 2015), tentou combinar a AS geral e orientada a alvos de uma determinada opinião em uma base de *tweets*. Os autores incluíram passos de pré-processamento como tokenização, tratamento de abreviações, gírias, negações, lematização, *Part-of-Speech*, reconhecimento de entidades, além de tratamento de *hashtags* e remoção de URLs. Por outro lado, um estudo desenvolvido por brasileiros (VITORIO; DOUGLAS, *et al.,* 2017) realizou uma análise considerando as principais diferenças na aplicação de AS para o PT_Br e o de Portugal. Para tal, os autores aplicaram módulos de NLP em *tweets* de ambos os países. Os mesmos notaram que treinando e testando *tweets* de diferentes variantes do português em um algoritmo de classificação resultava em uma piora no desempenho, comparando-se com o treinamento utilizando apenas uma variante. Assim, o estudo concluiu que ambas as variantes dessa língua podem causar impactos consideráveis na acurácia da AS.

3.3.2 Contexto Brasileiro

Em relação às pesquisas de MT e AS voltadas para a língua Portuguesa (PT), os mapeamentos sistemáticos realizados por Souza, *et al*. (2016a, 2016b), examinaram o

histórico de trabalhos apresentados na literatura e mostraram que 77% destes são voltados para o PT_BR, 62% tem o foco na aplicação de técnicas de AS e 60% são abordagens baseadas em dicionário. Ademais, o estudo ratificou a ausência de bases de dados públicas para replicação dos experimentos e assim avançar nas pesquisas. Para ratificar isto, o estudo sistemático de Sinoara, Antunes e Rezende (2017), afirma que somente 1% das pesquisas de MT apresentadas na literatura são voltadas para o PT e que há uma necessidade de construção ou expansão de recursos específicos para cada idioma.

A fim de se obter uma visão detalhada sobre a literatura de pré-processamento para AS, focado no português do Brasil, foi realizada uma revisão de literatura exclusiva nesta temática. Para esta revisão, adotou-se um processo sistemático de mapeamento similar a (SOUZA; BRUNO, *et al*., 2016) e inspirado por (KITCHENHAM; CHARTES, 2007) e (PETERSEN; KAI, *et al.,* 2008). A metodologia adotada é composta pelos seguintes passos: 1) definição das perguntas de pesquisa; 2) condução da busca; 3) seleção dos artigos com base nos critérios de inclusão; 4) seleção dos artigos com base nos critérios de qualidade; 5) extração de informações e mapeamento dos trabalhos. A revisão ficou em artigos escritos tanto em português quanto inglês, oriundos das seguintes bases: Science Direct, Scopus, Scielo, Capes e Google Scholar.

A partir desta revisão, foi possível se obter 61 artigos relevantes, no que tange a aplicação de pré-processamento para AS em PT_Br. Além disso, detectou-se quais as tarefas de pré-processamento mais adotadas para dados de mídias sociais do Brasil.

Presente em 46% dos estudos, a remoção de *stopwords* foi a tarefa mais mencionada. Em segundo, o tratamento de *hashtags*, citado em 36% das propostas. Estes passos foram seguidos por tokenização (34%), tratamento de menções (33%), e negação com tratamento de URLs, ambos em 31% dos estudos. Por outro lado, os métodos mais raramente encontrados foram o tratamento de letras em maiúsculas, repetições de letras, tratamento de sinais de exclamação e interrogação, risadas e cumprimentos. Cada um destes foi encontrado em apenas um estudo.

Referente aos conceitos descritos na Tabela 2.1, notou-se também uma variação quanto à metodologia de implementação para algumas tarefas de pré-processamento. Esta variedade foi detectada para as tarefas relacionadas ao tratamento de pontuações, repetições de letras, sinais de exclamação, *emoticons*, *hashtags*, menções e URLs. Estas variações incluem as ações de: Remoção, Colocação, Substituição e Computação de

Polaridade. A Tabela 4.1 exemplifica tais variações encontradas na literatura, quanto aos passos de pré-processamento supracitados, além das siglas já introduzidas no Capítulo 2 para identificar cada metodologia.

TABELA 3.1 VARIAÇÕES DE METODOLOGIA EM PRÉ-PROCESSAMENTO NO BRASIL.

Tarefa	Remoção (RM)	Colocação (CO)	Substituição (SB)	Computação de Polaridade (CP)
Sinais de Pontuação	X	X		
Repetições de Letras	X			X
Exclamação	X			X
Emoticons	X			X
Hashtags	X		X	X
Menções	X		X	
URLs	X		X	

No caso desta dissertação, a metodologia majoritariamente adotada para implementação nos passos de pré-processamento da arquitetura desenvolvida será a de Substituição por anotações ou identificadores. Este cenário será detalhado na Subseção 4.3.2.

No contexto brasileiro, alguns trabalhos se destacaram por sua relevância na literatura ou metodologias mais completas de pré-processamento. O primeiro foi (DURAN; AVANÇO; NUNES, 2015), um dos primeiros autores que propuseram um normalizador baseado em um dicionário léxico para conteúdo gerado por usuários para o PT_Br. Esta proposta era composta de passos como inserção automática de pontos, em seguida correção ortográfica fonética, tratamento de abreviações e gírias, além da identificação de entidades nomeadas através de nomes próprios. Todavia, estes autores não capturam qualquer tipo de sinal informal de intensificação ou negação de sentimento.

O trabalho de (STIILPEN JUNIOR; MERSCHMANN, 2016) apresentou uma metodologia para lidar com dados de mídias sociais, para fins de aplicações de Mineração de Texto, incluindo AS. Dentre suas estratégias, o autor incluiu correção de capitalização de palavras, correção ortográfica, tratamento de *hashtags* e menções, gírias e abreviações, *Part-of-Speech*, *stemming*, lematização e contextualização baseada em conceitos do site Wikipedia. A partir da métrica F1 com a adoção de um classificador SVM, os autores notaram o maior desempenho quando eles incluíram todos os passos de pré-processamento no seu *pipeline*. Porém, também não lidam com negações, amplificadores,

risadas ou palavras afetivas, os quais podem prejudicar o desempenho desta proposta quando presentes no texto.

O trabalho recente de (RODRIGUES; RAMON GOUVEIA, *et al*., 2016) também tentou lidar com conteúdo gerado por usuários, aplicando passos de pré-processamento citados anteriormente, porém para detectar o humor de pacientes em RSO. Em sua arquitetura, os autores incluíram tratamento de *emoticons*, *hashtags*, e intensificadores de sentimento. Por fim, os mesmos observaram as melhores médias para a métrica F1 com a adoção de tais passos para uma base de dados do Facebook composta por comentários positivos, negativos e neutros. Todavia, somente as tarefas mencionadas de pré-processamento são incorporadas, sem o tratamento de abreviações e gírias.

Outro recente destaque é o trabalho de (CORRÊA; EDILSON ANSELMO, *et al.,*2017). Neste, os autores tentaram realizar uma classificação com bases compostas de *tweets* de variados domínios. Os mesmos empregaram um total de dez tarefas de pré-processamento, incluindo o tratamento de intensificadores de sentimento e detratores. Este último foi um aspecto encontrado em apenas 8% da literatura existente. Como método de classificação, foi adotada a Supervisão de Distância, a fim de se construir uma base expandida de dados rotulados para AS. Treinando modelos de aprendizado de máquina em bases de mesmos e diferentes domínios, incluindo um *pipeline* de pré-processamento para conteúdo informal, os autores obtiveram resultados competitivos. Porém, sua proposta não lida com correção ortográfica e informalidades como abreviações e regionalismos, o que novamente pode afetar esta performance.

3.3.3 Síntese da Literatura

A Tabela 3.2 resume quais são as tarefas de pré-processamento executadas em todas as propostas discutidas nos contextos global e brasileiro. Para fins de uma melhor visualização, os artigos apresentados na tabela terão sua identificação de acordo com a seguinte lista de identificação:

1. (SINGH; KUMARI, 2016)
2. (MOSQUERA; GUTIÉRREZ; MOREDA, 2017)
3. (THELWALL; BUCKLEY; PALTOGLOU; CAI; KAPPAS, 2010)
4. (HUTTO; GILBERT, 2014)

5. (KHAN; BASHIR; QAMAR, 2014)
6. (FERSINI; MESSINA; POZZI, 2016)
7. (TELLEZ; ERIC, *et al.,*2017b)
8. (BALAHUR; PEREA-ORTEGA, 2015)
9. (TELLEZ; ERIC, *et al.,*2017a)
10. (MUSTO; SEMERARO; LOPS; GEMMIS, 2015)
11. (VILARES; THELWALL; ALONSO, 2015)
12. (SAIAS; JOSE, *et al.*, 2015)
13. (DURAN; AVANÇO; NUNES, 2015)
14. (STIILPEN JUNIOR; MERSCHMANN, 2016)
15. (RODRIGUES; RAMON GOUVEIA, *et al*., 2016)
16. (CORRÊA; EDILSON ALSEMO, *et al*., 2017)

TABELA 3.2 SÍNTESE DE TAREFAS DE PRÉ-PROCESSAMENTO APLICADAS NA LITERATURA.

Tarefa	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Tokenização				X		X	X	X				X	X	X		X
Stopwords	X				X		X		X				X			
Acentos									X					X		
Abreviações		X	X	X	X	X		X	X	X	X	X	X	X		
Gírias	X	X		X	X	X					X	X	X	X		
Emoticons		X	X	X	X	X	X	X	X		X		X	X	X	X
Correção ortográfica	X	X	X		X	X			X		X			X		
Correção Ortográfica Fonética		X											X			
Hashtag								X	X			X		X	X	X
Menções							X	X	X					X		X
URL							X		X			X	X	X		X
Negação			X	X			X	X	X		X	X				X
Repetições de Letras			X					X	X		X				X	
Intensificadores			X	X		X		X	X		X				X	X
Detratores			X	X				X			X					X
Maiúsculas				X											X	
Minúsculas	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Repetição de Exclamação								X								
Repetição de Interrogação								X								
Exclamação			X	X							X				X	
Interrogação			X												X	
Tratamento de Pontuação														X		
Remoção de Pontuação									X				X			
Números							X		X							
Risadas			X													
Palavras Afetivas								X								
Stemming							X	X	X	X				X		
Lematização					X				X			X		X		
PoS						X			X	X		X	X	X		
REN									X	X		X	X			
Context										X				X		
Total de Tarefas	**4**	**6**	**11**	**10**	**7**	**8**	**9**	**14**	**18**	**6**	**10**	**10**	**11**	**15**	**8**	**9**

A partir dos resultados, reforça-se a conclusão de que cada proposta busca implementar seu conjunto particular de passos de pré-processamento, variando em quais tarefas específicas e seu número de passos de tratamento de dados. A fim de se entender a razão para este comportamento, foi realizada a leitura minuciosa de cada artigo obtido a partir das revisões realizadas. Concluiu-se então que os autores, apesar de aplicação de pré-processamento, usualmente focam em suas bases de dados específicas. Assim, cada autor tem buscado as tarefas consideradas melhores para suas coleções e para se atingir a melhor precisão para as mesmas, sem a preocupação com uma generalização de pré-processamento para múltiplas bases.

Outro ponto observado foi que autores no contexto brasileiro não têm explorado certas variações de metodologias de pré-processamento reportadas na Tabela 2.1, para algumas tarefas. Por exemplo, um único autor brasileiro citou elementos como risadas e cumprimentos em dados de mídias sociais, as quais o mesmo considerou apenas para Remoção (GRANDIN; ADAN, 2016). Assim, ainda não foi realizada a investigação no cenário brasileiro de tais tarefas para, por exemplo, influenciar na computação da polaridade final de um texto.

Ainda comparando com propostas para outras línguas, elementos como sinais de exclamação e palavras afetivas foram substituídos por identificadores para um algoritmo de aprendizado de máquina em (GHIASSI; SKINNER; ZIMBRA, 2013) para a língua inglesa. Esta metodologia de Substituição foi adotada para o português somente no caso de *hashtags* e URLs por (STIILPEN JUNIOR; MERSCHMANN, 2016) e (CORRÊA; EDILSON ALSEMO, *et al*., 2017). Outro exemplo nesta direção é o trabalho de (TELLEZ; ERIC, *et al.,*2017b), com ênfase multilíngue, o qual adotou a ação de substituição para *emoticons*, atribuindo a polaridade atrelada aos mesmos como identificadores para classificadores de sentimento. Na literatura brasileira, *emoticons* têm sido adotados somente para a computação direta da polaridade, em propostas baseadas em dicionário léxico e de aprendizado de máquina (DE ARAUJO, GABRIELA DENISE, *et al*., 2018) e (PRATA, DAVID N., *et al.,* 2016).

Ainda objetivando um aprofundamento na razão pela não implementação de mais tarefas de pré-processamento nos trabalhos encontrados, o estudo de (SAIAS; JOSE, *et al*., 2015) reporta que a ausência de ferramentas de tratamento para a língua portuguesa,

na época de seu estudo, pode ser um motivo para tal comportamento. Em (DURAN; AVANÇO; NUNES, 2015), menciona-se o alto custo de manutenção no uso de listas que lidam com correção ortográfica fonética e demais tarefas de normalização para AS. Por fim, (TELLEZ; ERIC, *et al.,*2017a) destaca o alto consumo de memória e tempo de execução que pode ser ocasionado por um número elevado de tarefas de pré-processamento.

Quanto às bases de dados adotadas oriundas de RSO em PT_Br, bases do Twitter são a maioria, presentes em 70% dos estudos. O Facebook é menos adotado, sendo encontrado em apenas 4 estudos da literatura analisada. Demais bases presentes incluem revisões sobre produtos e serviços, além de comentários em *sites* de notícias.

Referente ao modelo de classificação para sentimento, propostas brasileiras têm adotado a estratégia estatística e de aprendizado de máquina em sua maioria (48% dos estudos). Nestas, os algoritmos mais utilizados foram o SVM e *Naive Bayes* (RISH, 2001). Dentre as demais técnicas mencionadas, tem-se *Multinomial Naive Bayes* (KIBRIYA, *et al*, 2004), Árvore de Decisão (APTÉ; WEISS, 1997), *Random Forest* (BREIMAN, 2001), Regressão Logística (HOSMER; LEMESHOW; STURDIVANT, 2013), *Ridge Classifier* (SAUNDERS, C., GAMMERMAN, A., & VOVK, 1998) e *Multilayer Perceptron* (GARDNER; DORLING, 1998). Particularmente, o algoritmo de SVM também tem sido adotado por diversas propostas no contexto global de AS (SUN; LUO;CHEN, 2017). A segunda metodologia de classificação mais adotada foi a afinidade léxica, em 26% dos estudos. Já o método de *keyword spotting* foi notado em 13% dos estudos. Estes números reforçam a preferência da literatura por modelos de aprendizado de máquina para trabalhar nesta temática.

Por fim, notou-se que não existe um trabalho que englobe todas as tarefas de pré-processamento encontradas na literatura de AS. Assim, ainda não houve a preocupação com a investigação do impacto de se expandir tais tarefas de tratamento de dados, além de diferentes estratégias de implementação de tais metodologias. Outro ponto observado foi que a maioria das tarefas aplicadas tem foco em ações de limpeza dos dados, onde cada autor implementa seu próprio conjunto de tarefas. Além disto, notou-se que esta dissertação está entre os trabalhos que aplicam o maior número de bases de dados para avaliação de desempenho de AS, totalizando 7, no contexto brasileiro.

Desta forma, pode-se confirmar que não há ainda um padrão uniforme de pré-processamento estabelecido para a classificação de sentimento em português do Brasil. Dado que bases de dados de mídias sociais apresentam diversos desafios, os quais são tratados de forma particular por cada uma destas tarefas, este trabalho se propõe a apresentar um *framework* a cobrir um maior número de tarefas de pré-processamento para AS em PT_Br, e consequentemente lidar com mais variantes de desafios no cenário de mídias sociais.

Assim, a Tabela 3.3 resume o diferencial da arquitetura SentiBR para a literatura existente brasileira.

TABELA 3.3 DIFERENCIAIS DO SENTIBR PARA A LITERATURA.

Literatura	**SentiBR**
Limitação em número de tarefas	Maior número de tarefas da literatura (23)
Não aplica de Substituição Risadas, Cumprimentos, Palavras Afetivas e emoticons	Primeira arquitetura a testar Substituição para estas tarefas no PT_Br
Pouca variação de bases de dados, com somente uma RSO ou revisões de produtos	Twitter, Facebook e notícias
Adoção de uma metodologia de classificação	Testes com múltiplas estratégias e seleção do melhor algoritmo classificador

3.3.4 Propostas para Comparação

Almejando realizar uma avaliação comparativa de desempenho da metodologia apresentada nesta dissertação, foram selecionadas três propostas do estado da arte.

O primeiro critério para a escolha das propostas foi que as mesmas tivessem sido desenvolvidas para uma linguagem de origem romântica, como é o caso do português. Tais exemplos incluem espanhol, italiano e francês. Estes idiomas compartilham origem similar e consequentemente têm aspectos em comum em seus núcleos linguísticos (PEETERS, 2006); e.g. o uso de acentos e gramática complexa. Além disso, outro simples exemplo que diferencia estas linguagens do inglês é o emprego da letra “a” no fim de adjetivos femininos, com a exceção do francês. Este fenômeno pode impactar o desempenho de ferramentas de NLP. Assim, a comparação com tais propostas nesse contexto, seria mais plausível e justa.

O segundo critério foi o número de passos executados de pré-processamento pelo *framework*, dado que se notou que algumas propostas implementam poucas tarefas, como (SINGH; KUMARI, 2016) e (MUSTO; SEMERARO; LOPS; GEMMIS, 2015). Por fim, o último critério foi o tipo e local de publicação das propostas. Optou-se por selecionar estudos submetidos a conferências e periódicos de alto impacto na literatura. Esta foi também a razão pela não adoção como *baselines* das propostas para o PT_Br presentes na Tabela 3.2.

Assim, a partir dos critérios supracitados, foram selecionadas para comparação as propostas de (BALAHUR; PEREA-ORTEGA, 2015), (TELLEZ; ERIC, *et al.,*2017a) e (VILARES; THELWALL; ALONSO, 2015), já mencionadas na Subseção 3.3.1.

A primeira proposta selecionada, (BALAHUR; PEREA-ORTEGA, 2015), implementa passos de pré-processamento focados em mídias sociais multilíngue, incluindo também linguagens românticas. Este modelo trabalha com a redução de letras repetidas, mas também considera este fenômeno como um sinal intensificador de sentimento. A proposta também substitui abreviações por suas formas expandidas. Quanto às demais tarefas de pré-processamento, este método adota a metodologia de Substituição. Por exemplo, ao se detectar uma palavra afetiva negativa ou positiva, a mesma é substituída por um identificador "NEGATIVE", "HNEGATIVE", "POSITIVE" ou "HPOSITIVE". No caso dos identificadores iniciados com "H", sua adoção é feita quando a palavra é muito positiva ou muito negativa, de acordo com a escala de sentimento utilizada na lista de palavras afetivas da implementação. Para a classificação de sentimento, o modelo emprega um SVM com otimização mínima sequencial e *kernel* linear.

A segunda proposta selecionada para comparação foi (TELLEZ; ERIC, *et al.,*2017a). Esta proposta realizou o estudo do impacto de transformações textuais em espanhol para a tarefa de AS. Uma versão preliminar deste *framework* também foi aplicada para bases em português em (TELLEZ; ERIC, *et al.,*2017b). O trabalho realiza a expansão de abreviações e correção gramatical por meio da redução de repetições de letras; e as demais tarefas são implementadas com a metodologia de Substituição por identificadores. Seu modelo de classificação aplica um SVM com *kernel* linear.

Por fim, o terceiro método para a análise comparativa foi (VILARES; THELWALL; ALONSO, 2015). Este algoritmo é inspirado no *SentiStrength*

(THELWALL, *et al.*, 2010), e foi desenvolvido para um estudo de caso de *tweets* sobre política em espanhol. Entretanto, os autores utilizaram a implementação original do *SentiStrength*, por meio de adaptações nos recursos utilizados pelo mesmo, para a língua espanhola. Tais recursos incluem listas com: intensificadores de sentimento, *emoticons*, palavras afetivas, palavras no padrão da norma culta, gírias, termos de ironia, negações, palavras usadas para se fazer perguntas e abreviações. Além destes, este modelo adota regras e heurísticas próprias para a implementação de cada tarefa de pré-processamento, incluindo sua correção ortográfica, e não provê a saída de cada tarefa de forma pública.

A Tabela 3.4 resume as informações referentes aos modelos selecionadas para comparação com esta proposta. Para a aplicação dos mesmos nos testes desta dissertação, buscou-se a implementação original quando disponível. Quando não, a reimplementação do método era realizada. Estes detalhes estão descritos na Subseção 4.4.2.

TABELA 3.4 PANORAMA DOS BASELINES DO ESTADO DA ARTE.

Autor	Título	Língua	Periódico	h-index
(TELLEZ; ERIC, *et al.*,2017a)	A case study of Spanish text transformations for twitter sentiment analysis	Espanhol	Expert Systems with Applications	92-122
(BALAHUR; PEREA-ORTEGA, 2015)	Sentiment analysis system adaptation for multilingual processing: The case of tweets	Multilíngue	Information Processing & Management	35-48
(VILARES; THELWALL; ALONSO, 2015)	The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets	Espanhol	Journal of Information Science	22-28

3.4 CONSIDERAÇÕES FINAIS

Neste capítulo apresentaram-se os principais trabalhos que deram suporte e motivaram esta dissertação. Inicialmente, foram discutidos os desafios e conceitos relacionados a tarefa de pré-processamento em Análise de Sentimento. As tarefas adotadas e seu funcionamento também foram ilustrados. Neste ensejo, percebeu-se a carência por uma metodologia genérica para pré-processamento, que incorpore um conjunto expandido de tarefas de tratamento de dados para o cenário de mídias sociais.

Foi possível também, por meio da análise comparativa entre propostas de destaque na literatura, identificar a falta de padronização, tanto no contexto universal quanto brasileiro, abrangendo esta etapa crucial antes da classificação de sentimento. Este cenário torna notória a necessidade de uma averiguação do impacto da padronização de pré-processamento, além da possibilidade de uma nova arquitetura para esta tarefa em português do Brasil; para tal, considerando etapas que lidam com desafios variados no conteúdo gerado por usuários e assim propiciar uma comparação fidedigna entre as pesquisas desenvolvidas na área.

Estas lacunas identificadas na literatura serviram de subsídio para o desenvolvimento da arquitetura SentiBR nesta dissertação, a qual é apresentada com maiores detalhes no próximo capítulo.

4 ARQUITETURA DE PRÉ-PROCESSAMENTO PARA ANÁLISE DE SENTIMENTO EM PORTUGUÊS BRASILEIRO

4.1 CONSIDERAÇÕES INICIAIS

Como discutido nos capítulos anteriores, a problemática de falta de padronização de métodos de pré-processamento de dados de mídias sociais pode afetar negativamente a aplicação de AS. A falta de padronização nos experimentos, envolvendo diferentes tarefas de tratamento de dados por parte dos estudos no contexto brasileiro, dificulta a comparação entre os diversos métodos propostos e também a adoção dos novos métodos neste domínio.

Assim, a pesquisa desta dissertação foi desenvolvida a partir da seguinte hipótese: a adoção de uma padronização na etapa de pré-processamento de dados de mídias sociais aumenta a qualidade da tarefa de AS, propiciando a plena incorporação de indicadores de sentimento.

Este capítulo também apresenta, a metodologia e os processos desenvolvidos para executar a implementação da arquitetura de pré-processamento padronizada, SentiBR. Inicialmente, será ilustrada a agenda de pesquisa adotada para a condução do projeto. Em seguida, os componentes da arquitetura desenvolvida serão detalhados. Por fim, as ferramentas e detalhes de implementação serão ressaltados.

4.2 METODOLOGIA

Para a condução do projeto, foi adotada uma agenda de pesquisa focada em estágios chave para atingir as metas de desenvolvimento estabelecidas. A Figura 4.1 ilustra esta agenda.

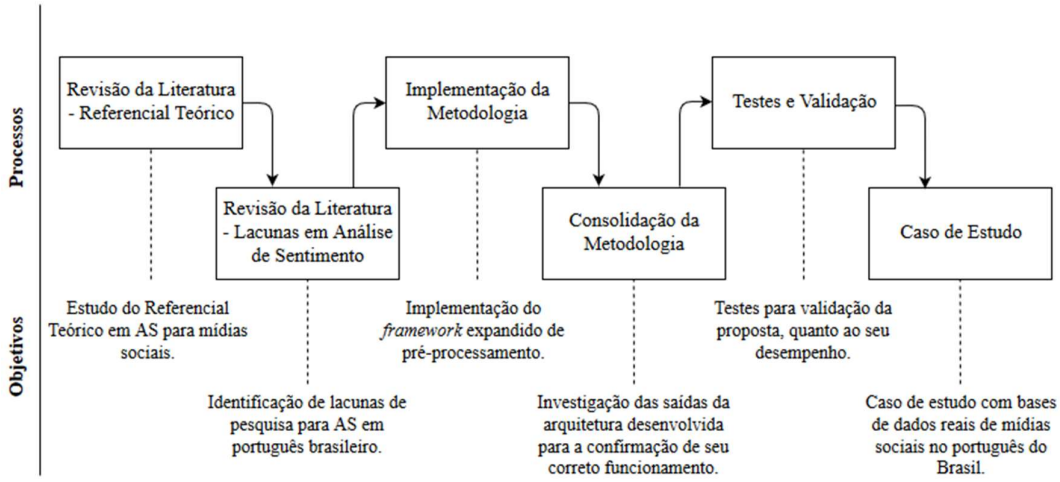

Figura 4.1 Agenda de pesquisa adotada para condução do estudo.

- **Revisão de Literatura – Referencial Teórico:** Durante esta etapa, o referencial teórico em AS e desafios de mídias sociais foram investigados. Os resultados desta etapa estão detalhados nos capítulos 2 e 3.
- **Revisão de Literatura – Lacunas em Análise de Sentimento:** Neste estágio, objetivou-se detectar lacunas na literatura em AS para o PT_Br, além de propostas existentes para o cenário de mídias sociais. Os resultados deste estágio estão detalhados no capítulo 3, de trabalhos correlatos.
- **Implementação da Metodologia:** A partir das lacunas e oportunidades de contribuição detectadas previamente, os esforços foram direcionados para a implementação de uma proposta que apresentasse contribuição na literatura investigada. Os detalhes deste estágio estão presentes neste capítulo.
- **Consolidação da Metodologia:** Neste momento, os primeiros testes foram realizados, a fim de se verificar potenciais erros e *bugs* a serem corrigidos para uma posterior rodada de validação do *framework*. Além disso, verificou-se se as saídas de cada tarefa de pré-processamento da arquitetura estavam sendo dadas como esperado.
- **Testes e Validação:** Algumas das bases de dados selecionadas para realização dos experimentos foram dadas como entrada para a arquitetura desenvolvida. Neste estágio, o ênfase foi em verificar se o desempenho estava equiparável ao estado da arte.

- **Casos de Estudo:** Com a arquitetura validada, prosseguiu-se para a realização dos casos de estudo, com todas as bases de dados selecionadas para o *benchmarking* comparativo da arquitetura, com os demais *baselines* selecionados para o projeto. Por fim, os resultados finais quanto ao desempenho da proposta foram obtidos e serão apresentados no capítulo 5.

4.3 ARQUITETURA

Para o desenvolvimento do *framework* proposto, adotou-se uma metodologia de implementação modular. O objetivo foi permitir que cada componente da mesma pudesse ser atualizado com menos complexidade.

4.3.1 Componentes da Arquitetura

A Figura 4.2 ilustra a visão geral da arquitetura desenvolvida com seu módulo principal de pré-processamento.

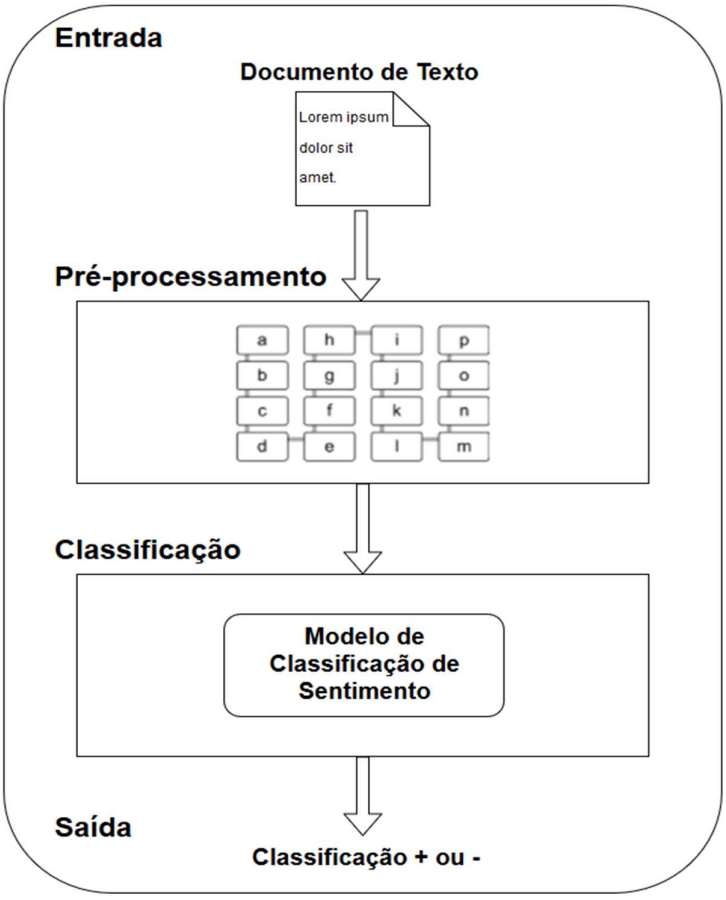

Figura 4.2 Visão geral da arquitetura proposta nesta dissertação.

O *framework* desenvolvido contém todas as tarefas de pré-processamento. O mesmo é responsável pela limpeza e formatação dos dados textuais e tem como saída os dados pré-processados para serem utilizados como entrada para um modelo de classificação adotado para a AS. Para o presente estudo, foi adotado um modelo de aprendizado de máquina SVM.

4.3.2 Tarefas de Pré-processamento

A arquitetura SentiBR proposta tem como objetivo ser uma plataforma expandida de pré-processamento para AS. Dado que os trabalhos correlatos encontrados implementam seu próprio conjunto de tarefas de pré-processamento, esta metodologia objetiva cobrir a maioria das mesmas, que são focadas em dados de mídias sociais. Assim, um total de 23 tarefas de pré-processamento são incorporadas neste *framework*.

A Figura 4.3 ilustra a visão mais detalhada sobre todas as tarefas de pré-processamento aplicadas no *framework*. A imagem apresenta 20 módulos de tratamento, porém alguns dos mesmos incorporam mais de uma tarefa de pré-processamento ou variação da mesma. Por exemplo, o módulo de amplificadores e detratores emprega 5 variações de intensificação de sentimento, além de advérbios de intensificação. Desta forma, o total de tarefas empregadas permanece em 23. A razão para a implementação neste formato, é para adotar uma compactação de tarefas, e assim facilitar a manutenção de código por futuros pesquisadores na área. Com isso, para tratar de tarefas que envolvem amplificação de sentimento, por exemplo, basta trabalhar em um único módulo da arquitetura.

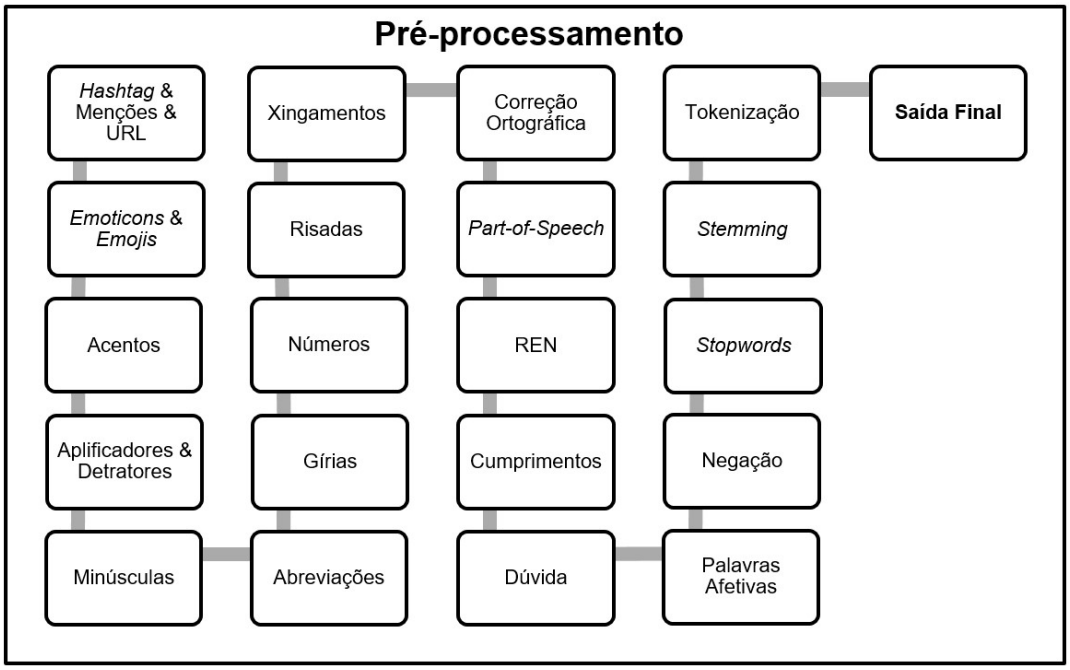

Figura 4.3 Tarefas de pré-processamento empregadas na arquitetura deste estudo.

As tarefas de pré-processamento são organizadas de forma sequencial, de modo que o fluxo de informações e saídas de cada tarefa é repassado ao próximo, para seu respectivo processamento sobre os dados. Usuários da plataforma estarão aptos a desligar módulos específicos do *pipeline* de pré-processamento. Entretanto, a acurácia da análise pode variar de acordo com as tarefas a serem aplicadas.

A forma principal de operação da maioria dos métodos de pré-processamento segue a metodologia de substituição, descrita na Subseção 2.3.2. Assim, as mesmas objetivam prover anotações sobre os dados, a fim de fornecer um conjunto enriquecido de informações para um classificador de sentimento. Com isto, algumas tarefas provêm uma anotação específica, que corresponde à característica encontrada no texto, que pode indicar um potencial sinal de um dado sentimento. Por outro lado, outras tarefas realizam transformações nos dados sem prover anotação, esse é o caso da tokenização, conversão para minúsculas, remoção de acentos, remoção de *stopwords*, e *stemming*.

As anotações e transformações aplicadas pela arquitetura tem como objetivo capturar e normalizar sinais indicadores de informação afetiva, polaridade ou sentimento presente em um texto. Quando por exemplo, uma palavra negativa, como “horrível” é

encontrada, a arquitetura converte este *token* específico para a anotação “negative”, o que significa que um sinal de sentimento negativo foi encontrado naquela parte do texto.

Conforme mencionado anteriormente, existem variações de um mesmo sinal capturado por uma tarefa de pré-processamento. Como exemplo, um *emoticon* ou *emoji* negativo e uma palavra negativa têm as anotações “neg_emotj” e “negative” atribuídas. Porém, para fins de normalização, o *framework* atribui a anotação “negative” para ambos os casos. Assim, para algumas tarefas, existem dois tipos de anotações disponíveis na arquitetura: i) o primeiro é a anotação detalhada, a qual provê informação sobre o tipo específico de aspecto encontrado (“neg_emotj” ou “pos_emotj”); ii) o segundo tipo é a anotação geral, a qual fornece informação quanto ao sentimento geral de um dado aspecto no texto (“positive” e “negative”).

Para alguns aspectos, a anotação detalhada foi mantida, como é o caso de xingamentos (“dirtyword”), cumprimentos (“greeting”), números (“numbers”), padrões de risos (“laughing”), advérbios de dúvida (“doubt”), e partículas de negação (“negation”). A razão para essa configuração, é que em todos esses casos não se pode atribuir uma conotação negativa ou positiva com alta acurácia. Por exemplo, cumprimentos podem ser encontrados tanto em textos positivos quanto negativos, assim como risos e números.

Com todas essas anotações, o objetivo principal é que o classificador de sentimento consiga aprender a identificar a polaridade de um texto a partir da presença de tais identificadores. Cada identificador terá sua polaridade aprendida a partir dos seus pesos correspondentes, dados como entrada para um modelo computacional a partir da representação mencionada na Subseção 2.2.1.4.

A Tabela 4.1 apresenta o esquemático de como cada tarefa de pré-processamento é implementada na arquitetura proposta. O *framework* também permite variações quanto à forma de operação e parâmetros, de acordo com as metodologias apresentadas na Tabela 2.1. As anotações fornecidas por cada tarefa, quando realizadas, também são mencionadas.

TABELA 4.1 PANORAMA DAS TAREFAS DE PRÉ-PROCESSAMENTO DA ARQUITETURA.

Tarefa	Metodologia	Variações	Anotação Detalhada	Anotação Geral	Exemplo de Entrada	Exemplo de Saída
Hashtags	SB	Substituição de *hashtag* por identificador	hashtag	hashtag	O #BancoDoBrasil me deu parabéns hoje!	O hashtag me deu parabéns hoje!
	SB	Remoção do *character* "#" e manutenção do conteúdo da *hashtag*.	hashtag	hashtag	O #BancoDoBrasil me deu parabéns hoje!	O BancoDoBrasil me deu parabéns hoje!
	TR	Tratamento de *Hashtags* pela separação de termos compostos.	hashtag	hashtag	O #BancoDoBrasil me deu parabéns hoje!	O Banco Do Brasil me deu parabéns hoje!
Menções	SB	-	mention	mention	A @Globo poderia passar aquele filme né!	A mention poderia passar aquele filme né!
URL	SB	-	url	url	Olha esse link!: http://link.com.br	Olha esse link!: url
Emoticons* e *Emojis	SB	-	pos_emotj ou neg_emotj	positive ou negative	Uau! ! Quanta alegria e tristeza : (	Uau! positive! Quanta alegria e tristeza negative
Acentos	TR	-	-	-	Eu AMO aquele filmeee incrível!!!	Eu AMO aquele filmeee incrivel!!!
Amplificadores	SB	Maíusculas	upperintens	intensifier	Eu AMO aquele filmeee incrível!!!	Eu intensifier aquele filmeee incrível!!!
	SB	Repetições de Letras	lenintens	intensifier	Eu AMO aquele filmeee incrível!!!	Eu AMO aquele intensifer incrível!!!
	SB	Repetições de Exclamação	excintens	intensifier	Eu AMO aquele filmeee incrível!!!	Eu AMO aquele filmeee incrível intensifier
	SB	Repetições de Interrogação	questionintens	intensifier	E você ainda pergunta????	E você ainda pergunta intensifier
	SB	Advérbios de Intensidade como Amplificadores	ampintens	intensifier	Eu quero muito isso!	Eu quero intensifier isso!

TABELA 4.1 PANORAMA DAS TAREFAS DE PRÉ-PROCESSAMENTO DA ARQUITETURA. (CONT.)

Tarefa	Metodologia	Variações	Anotação Detalhada	Anotação Geral	Exemplo de Entrada	Exemplo de Saída
Detratores	SB	-	downintens	downtoner	Acho pouco molho	Acho downtoner molho
Minúsculas	TR	-	-	-	Eu AMO aquele filmeee incrível!!!	eu amo aquele filmeee incrível!!!
Abreviações	CR	-	abbreviation	abbreviation	Acho aquele livro é tdb!	Acho aquele livro é tudo de bom!
Gírias	SB	-	pos_slang ou neg_slang	positive ou negative	Aquela praça está largada às traças!	Aquela praça está negative!
Números	SB	-	number	number	Eu comprei 7 livros ontem!	Eu comprei number livros ontem!
Risadas	SB	-	laugh	laugh	hahahahaha que show!	laugh que show!
Xingamentos	SB	-	dirtyword	dirtyword	Achei uma merda esse time!	Achei uma dirtyword esse time!
Correção Ortográfica	CR	-	spell_error	spell_error	Foi erada aquela atitude!	Foi errada aquela atitude!
Part-of-speech	CP	-	´part_postag	part_ postag	Eu quero muito isso!	Eu (PRN), quero (VB), muito (ADV), isso (PRN), ! (PUNC)
Reconhecimento de Entidades	CP	-	ner	ner	Presidente Obama nos enviou uma mensagem!	Presidente Obama (NER) nos enviou uma mensagem!
Cumprimentos	SB	-	greeting	greeting	Bom dia! O que temos hoje?	greeting! O que temos hoje?
Dúvida	SB	-	doubt	doubt	Talvez você goste!	doubt você goste!
Palavras Afetivas	SB	-	pos_word ou neg_word	positive ou negative	Eu curto aquele filmeee incrível, mas o protagonista é ruim!!!	Eu positive aquele filmeee positive, mas o protagonista é negative!!!

TABELA 4.1 PANORAMA DAS TAREFAS DE PRÉ-PROCESSAMENTO DA ARQUITETURA. (CONT.)

Tarefa	Metodologia	Variações	Anotação Detalhada	Anotação Geral	Exemplo de Entrada	Exemplo de Saída
Negação	SB	Ingênua: inverte o sentimento de termos dentro de uma janela de *tokens*	negation	negation	O número final não saiu tão bem como nos belos ensaios	O número final negation saiu tão negative como nos negative ensaios
	SB	Padrão: inverte o sentimento de termos dentro de uma janela de *tokens*, caso a partícula de negação anteceda um verbo	negation	negation	Eu não gostei do sorvete!	Eu negation negative do sorvete
Stopwords	RM	-	-	-	eu amo aquele filmeee incrível!!!	amo filmeee incrível!!!
Stemming	TR	-	-	-	eu amo aquele filmeee incrível!!!	eu am aquel film incrível!!!
Tokenização	TR	-	-	-	Eu AMO aquele filmeee incrível!!!	Eu, AMO, aquele, filmeee, incrível, !!!

4.3.3 Saídas da Arquitetura de Pré-processamento

Após todas as tarefas de pré-processamento serem aplicadas em um documento, seu formato de saída irá conter as anotações e transformações apresentadas na Subseção 4.3.2. A Tabela 4.2 a seguir ilustra o impacto de cada método em um texto de entrada, até a sua saída final para a etapa de classificação. Para um melhor comparativo com o estado da arte, a Tabela 4.2 também apresenta as saídas dos *baselines* a serem empregados no capítulo 5, para a mesma entrada textual. A saída do modelo (VILARES; THELWALL; ALONSO, 2015), o qual também será um *baseline*, não é ilustrada devido a implementação e código deste não estarem disponíveis de forma pública.

O texto de entrada para todos os métodos é: "kkkkk pqp! Eu AMEI esse novo modelo!!! Mas talvez ainda possa melhorar".

TABELA 4.2 EXEMPLO DE SAÍDA DA ARQUITETURA E *BASELINES*.

Modelo	Saída Final
SentiBR	laugh dirtyword positive nov model intensifier doubt po positive
(BALAHUR; PEREA-ORTEGA, 2015)	kkkkk pqp eu positive esse novo modelo multiexclamation mas talvez ainda pode positive
(TELLEZ; ERIC, *et al.*, 2017a)	kkkkk pqp ame nov model talvez po melhor

A partir da Tabela 4.2, nota-se que o SentiBR foi capaz de capturar o maior número de elementos, que possam indicar sinais de sentimento no texto de entrada. Por exemplo, enquanto os *baselines* não conseguiram prover anotações para a presença de risadas, xingamento e advérbio de dúvida, estes sinais foram capturados e normalizados pela proposta. Demais sinais, tais como a repetição de exclamação como intensificação e palavras afetivas podem ser detectados pela proposta e por (BALAHUR; PEREA-ORTEGA, 2015). Entretanto, estes não são sinais capturados por (TELLEZ; ERIC, *et al.*, 2017a), o que pode se caracterizar como uma perda de potencial informação afetiva para este modelo, dado que palavras afetivas e sinais de amplificação podem ser importantes indicadores de polaridade em um texto.

4.3.4 Classificação de Sentimento

Para esta dissertação foi escolhido o algoritmo SVM para executar a classificação do sentimento. Este foi configurado com o *kernel* linear. A razão para a escolha da técnica SVM deu-se após investigação da literatura de aprendizado de máquina para AS. Notou-se que o SVM é um dos algoritmos mais utilizados, além de apresentar maior desempenho em comparação com demais modelos de aprendizado para esta tarefa (SOUZA, ELLEN, *et al*, 2016).

Como apresentado anteriormente no capítulo 2, modelos computacionais de classificação de sentimento necessitam de uma representação numérica dos dados textuais. Para este classificador na arquitetura proposta, adotou-se o modelo descrito na Subseção 2.2.1.4, denominado *Bag-of-Words*. O mesmo foi aplicado com o peso TF-IDF para a matriz de Documentos-Termos.

Dessa forma, o classificador SVM recebe como entrada uma matriz de Documentos-Termos, onde cada linha contém o vetor de atributos para cada documento da base de dados e cada coluna o peso TF-IDF para cada termo único do vocabulário.

4.3.5 Configurações do *Framework*

O *framework* possui parâmetros que podem ser configurados para sua execução. Os mesmos são escolhidos a partir de um arquivo XML[19], que contém os valores possíveis para cada parâmetro. A Tabela 4.3 ilustra todas essas opções.

TABELA 4.3 PARÂMETROS PARA CONFIGURAÇÃO DO *FRAMEWORK* DE PRÉ-PROCESSAMENTO.

Parâmetro	Opções	Efeito
dataTransformationWeight	• TP • TF • TF-IDF	O peso a ser utilizado para a etapa de transformação de dados, e composição da matriz Documentos-Termos a ser dada como entrada para o classificador de sentimento.
nfolds	• Valor inteiro a partir de 2	O número de rodadas a ser realizada na validação cruzada para avaliação de AS.

[19] https://www.w3.org/2000/xp/

TABELA 4.3 PARÂMETROS PARA CONFIGURAÇÃO DO *FRAMEWORK* DE PRÉ-PROCESSAMENTO. (CONT.)

Parâmetro	Opções	Efeito
method_ HashtagMentionUrlLight	• Substituição • Substituição e manutenção de conteúdo • Tratamento	Configura o modo de tratamento de *hashtags*¸menções e URLs
posTagger	• Polyglot • NLPNet	Seleciona a ferramenta a ser utilizada para o processo de *Part-of-Speech tagging*.
nerTool	• Polyglot • Spacy	Seleciona a ferramenta a ser utilizada para o processo de Reconhecimento de Entidades Nomeadas.
doubt_POSCheck	• Verdadeiro • Falso	Define se os termos considerados como advérbios de dúvida realmente receberam a *tag* de advérbio pelo processo de *Part-of-Speech*.
method_sentimentWords	• Simples • Múltiplo	Define se a escala de anotações para palavras afetivas será Simples (positiva ou negativa), ou Múltipla (incluindo as anotações muito positiva e muito negativa).
method_negation	• Ingênuo • Padrão	Define forma de operação da tarefa de negação. O modelo ingênuo permite a inversão de polaridade de todos os *tokens* dentro de uma janela definida. A opção padrão só realiza esta inversão se a partícula de negação vier antes de um advérbio no texto.

4.4 DESENVOLVIMENTO

Esta subseção se concentra em apresentar os pacotes e bibliotecas de *software*¸ bem como ferramentas utilizadas para o desenvolvimento da arquitetura neste trabalho.

4.4.1 Implementação do *Framework*

A linguagem de programação majoritariamente utilizada para o desenvolvimento deste *framework* foi o Python, em sua versão 2.7. O ambiente de desenvolvimento escolhido foi o *software* Sublime Text, em sua versão 3. O sistema operacional para desenvolvimento foi a distribuição Linux chamada Ubuntu, em sua versão 16.04 LTS.

A razão para a adoção da linguagem Python é que a mesma possui uma comunidade de desenvolvimento bastante ativa, o que permite as frequentes atualizações de seus recursos. Além disso, são diversas as bibliotecas desta linguagem voltadas para aprendizado de máquina e mineração de dados, e as comunidades de cientista de dados a utilizam constantemente. Quanto ao *software* Sublime Text, o mesmo apresenta uma interface amigável e completa para desenvolvimento, também com a possibilidade de instalação de *plugins* e alterações para a melhor experiência de desenvolvimento por parte de seus usuários.

Uma biblioteca de grande importância para o funcionamento desta arquitetura é a responsável pelo processamento de expressões regulares. Estas são a base do pré-processamento desta ferramenta. Por exemplo, para a captura dos aspectos descritos na tabela 4.1, são utilizadas *regex* para a busca de tais sinais no texto. A exceção é para as tarefas de *Part-of-Speech,* Reconhecimento de Entidades, Correção Ortográfica, *Stemming*, Acentos e Minúsculas, que possuem sua própria forma de operação.

Optou-se por este modo de operação na tentativa de redução do tempo de execução de tais tarefas de pré-processamento, evitando a adoção de laços sobre os *tokens* presentes no texto. Um exemplo de *regex* adotada para a detecção de palavras afetivas é: "**b**amor**b**|**b**gosto**b**|**b**bom**b**". Neste exemplo, as palavras "amor", "gosto" e "bom" podem ser capturadas em um texto, desde que as mesmas estejam no início, fim ou entre espaços, o que é configurado pelo elemento "**b**". Isto evita confusões do tipo capturar o termo "amor" na palavra "amora" e assim processar um falso positivo para uma palavra afetiva.

Assim, a Tabela 4.4 introduz as bibliotecas e ferramentas utilizadas para o desenvolvimento deste *framework*.

TABELA 4.4 BIBLIOTECAS E FERRAMENTAS PARA O DESENVOLVIMENTO DO *FRAMEWORK*.

Biblioteca	Descrição
Scikit-learn[20]	Biblioteca com diversos algoritmos de aprendizado de máquina para o Python.
NLTK	Biblioteca com métodos e recursos de NLP para o Python.
Pandas[21]	Biblioteca *open source*, que fornece estruturas de dados de alto desempenho e ferramentas de análise de dados para o Python.
Numpy	Biblioteca para computação científica no Python. Por exemplo, possui funções para se trabalhar com álgebra linear e números aleatórios.
SnowballStemmer[22]	Ferramenta de *Stemming* disponível no NLTK. Funciona também para a língua portuguesa.
Pyfreeling[23]	Biblioteca para aplicação da ferramenta Freeling em Python, a qual é executa a tarefa de lematização.
BeautifulSoup	Biblioteca para extrair dados de arquivos HTML e XML para o Python.
lxml[24]	Biblioteca para processamento de arquivos XML e HTML para o Python.
Spacy	Biblioteca com métodos e recursos de NLP para o Python.
Polyglot[25]	Biblioteca que possui um *pipeline* de NLP para múltiplas linguagens, incluindo o português.
Hunspell[26]	Verificador ortográfico do LibreOffice, OpenOffice.org, Mozilla Firefox 3 e Thunderbird. Este é empregado na tarefa de correção ortográfica da arquitetura.
language_check[27]	Biblioteca Python para a aplicação da ferramenta LanguageTool. A mesma executa tarefas de correção ortográfica.
python-Levenshtein[28]	Biblioteca Python para calcular distâncias e semelhanças entre objetos do tipo *string*.

[20] http://scikit-learn.org/
[21] https://pandas.pydata.org/
[22] http://www.nltk.org/howto/stem.html
[23] https://github.com/malev/pyfreeling
[24] https://lxml.de/
[25] http://polyglot.readthedocs.io/en/latest/index.html
[26] http://hunspell.github.io/
[27] https://pypi.org/project/language-check/
[28] https://pypi.org/project/python-Levenshtein/

TABELA 4.4 BIBLIOTECAS E FERRAMENTAS PARA O DESENVOLVIMENTO DO *FRAMEWORK*. (CONT.)

Biblioteca	Descrição
re[29]	Biblioteca Python para aplicação de expressões regulares.
csv[30]	Biblioteca Python para manipulação de arquivos no formato *Comma-Separated Values*, ou csv.
UGCNormal	Ferramenta de normalização de dados de mídias sociais, incluindo o processo de correção ortográfica fonética. A mesma é originada do estudo de (DURAN; AVANÇO; NUNES, 2015).

Quanto aos demais recursos empregados para as tarefas de identificação de *emoticons, emojis,* abreviações, gírias, xingamentos, amplificadores e detratores de sentimento, advérbios de dúvida, partículas de negação e palavras afetivas, a Tabela 4.5 ilustra as respectivas fontes.

TABELA 4.5 RECURSOS ADOTADOS PARA IMPLEMENTAÇÃO DE TAREFAS DE PRÉ-PROCESSAMENTO.

Tarefa	Recurso
Emoticons	Lista[31] com *emoticons* comumente encontrados em textos na Internet.
Emojis	Lista [32] com *emojis* comumente encontrados em textos na Internet, e aplicativos de troca de mensagens. Este recurso provê ainda o nível de sentimento positivo ou negativo para cada *emoji*.
Abreviações	Lista de abreviações em português do Brasil, e seus respectivos correspondentes expandidos (GONZALEZ, 2007). Algumas abreviações mais recentes também foram adicionadas manualmente nesta lista.
Gírias	Lista com gírias e expressões idiomáticas usuais no português do Brasil (RIVA, 2009). Algumas gírias mais recentes foram adicionadas manualmente neste recurso.
Xingamentos	Lista[33] com 397 palavrões e expressões ofensivas para bloquear em canais de mídias sociais, e que se caracterizam como xingamentos. Alguns termos também foram adicionados manualmente nesta tarefa.

[29] https://docs.python.org/2/library/re.html
[30] https://docs.python.org/2/library/csv.html
[31] https://en.wikipedia.org/wiki/List_of_emoticons
[32] http://kt.ijs.si/data/Emoji_sentiment_ranking/index.html
[33] http://www.brunoavila.com/palavroes-especial-eleicoes

TABELA 4.5 RECURSOS ADOTADOS PARA IMPLEMENTAÇÃO DE TAREFAS DE PRÉ-PROCESSAMENTO. (CONT.)

Tarefa	Recurso
Amplificadores & Detratores	Para a aplicação desta tarefa, foram adotadas múltiplas listas contendo termos e advérbios de intensidade e redução de sentimento. A primeira é composta de sinônimos das palavras "muito" e "pouco", obtida de um site[34] especializado em sinônimos do português. As demais listas vem dos trabalhos de (AVANÇO, 2015), (FOLTRAN; NOBREGA, 2016) e (GONÇALVES, 2002).
Advérbios de Dúvida	Lista de advérbios de dúvida obtida de um portal[35] especializado na gramática da língua portuguêsa.
Negação	Lista de negação obtida a partir da tradução da mesma lista de (VILARES; THELWALL; ALONSO, 2015).
Palavras Afetivas	Lista com palavras e suas respectivas polaridades, obtida de (VILARES; THELWALL; ALONSO, 2015). A mesma também é empregada pelo *baseline* (BALAHUR; PEREA-ORTEGA, 2015).

A Figura 4.4 ilustra uma parte da lista de *emojis* adotada para a tarefa de pré-pré-processamento adotada para lidar com este elemento. Notam-se dois exemplos de *emojis*, e na coluna denominada *"Sentiment bar"*, tem-se uma escala de sentimento para cada polaridade, seja positiva, negativa ou neutra. A partir desta escala, foi possível se rotular os *emojis* obtidos desta lista.

Emoji Sentiment Ranking v1.0

Char	Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)	Unicode name
😂		0x1f602	14622	0.805	0.247	0.285	0.468	0.221		FACE WITH TEARS OF JOY
❤		0x2764	8050	0.747	0.044	0.166	0.790	0.746		HEAVY BLACK HEART

Figura 4.4. *Emoji Sentiment Ranking*: lista de *emojis* adotada na arquitetura.

[34] https://www.sinonimos.com.br/
[35] https://www.soportugues.com.br/

4.4.2 Implementações do Estado da Arte

A implementação de cada *baseline* será detalhada nesta seção.

- **(TELLEZ; ERIC, *et al*.,2017a)**

O primeiro a ser destacado é o modelo de (TELLEZ; ERIC, *et al*., 2017a). Para a utilização desta proposta, inicialmente tentou-se contato com os autores da mesma, a fim de se obter o código utilizado em seu artigo publicado.

Os autores gentilmente proveram a ferramenta. Entretanto, a mesma foi desenvolvida para um cenário multilíngue, incluindo línguas românticas como o espanhol e italiano. Como o português não estava incluindo, os recursos a serem utilizados para esta língua foram desenvolvidos. Além disso, seguindo o padrão da implementação para o espanhol, os métodos de pré-processamento foram replicados para a língua portuguesa.

Quanto aos recursos utilizados por esta proposta, incluem-se a ferramenta FreeLing[36] para a lematização. Esta também tem versão disponível para o português, a qual foi aplicada na adaptação do modelo para esta linguagem. Este autor também emprega tarefas que envolvem listas de negações, *stopwords* e palavras na norma culta para correção ortográfica. Para implementar tais listas correspondentes para a adaptação desta proposta em português, buscou-se por versões desses recursos mencionados para a língua portuguesa. Quando estas não eram encontradas, os mesmos recursos aplicados pela arquitetura desenvolvida nesta dissertação eram incorporados. Assim, foi possível reproduzir fielmente este trabalho, para a experimentação realizada neste projeto.

- **(BALAHUR; PEREA-ORTEGA, 2015)**

No caso do segundo *baseline*, também se tentou contato com os autores para se obter a sua implementação. Todavia, neste caso não se obteve um retorno dos mesmos. Desse modo, o artigo de (BALAHUR; PEREA-ORTEGA, 2015) foi implementado na íntegra, para uso como modelo comparativo.

Assim, seguindo o padrão já adotado nas demais implementações deste projeto, optou-se pelo desenvolvimento deste modelo na linguagem Python. Todos os passos de

[36] http://nlp.lsi.upc.edu/freeling/index.php/node/1

pré-processamento mencionados em seu artigo foram reproduzidos fielmente na implementação.

As anotações aplicadas por este modelo, também seguiram o mesmo padrão indicado em seu estudo. Quanto às ferramentas utilizadas pelo mesmo, observou-se a necessidade da inserção de listas em português para a tarefa de correção ortográfica, abreviações, amplificadores & detratores de sentimento, negações e palavras afetivas. Nestes casos, buscou-se aplicar a mesma lista mencionada no artigo original de (BALAHUR; PEREA-ORTEGA, 2015), porém em uma versão em português, caso disponível. Um exemplo disto era a ferramenta adotada pelos autores para a tarefa de lematização, chamada de *TreeTagger*[37]. A mesma tem a sua versão em português, de modo que esta foi utilizada na implementação deste *baseline.*

Quando os mesmos recursos não estavam acessíveis para o português, a mesma lista aplicada pela proposta desta dissertação era utilizada. Estas medidas foram adotadas, para se ter uma comparação justa entre os modelos concorrentes na experimentação final. Deste modo, foi novamente possível se obter uma representação fiel deste método do estado da arte, para a avaliação de desempenho.

- **(VILARES; THELWALL; ALONSO, 2015)**

O último *baseline* obtido foi o *SentiStrength* em espanhol, que foi convertido para uma versão brasileira para utilização no projeto dos referidos autores. Neste caso, novamente tentou-se contato com os autores da proposta. Os mesmos foram gentis em fornecer a implementação completa. Esta é composta de um arquivo .jar, e demais arquivos auxiliares para a operação do modelo;

Tais arquivos são compostos de recursos linguísticos, tais como listas de intensificadores de sentimento, *emoticons*, palavras afetivas, palavras no padrão da norma culta, gírias, termos de ironia, negações, palavras para pergunta e abreviações.

A orientação recebida dos autores do modelo, para a construção da versão em Português, foi para substituir todas as listas mencionadas por versões na língua portuguesa. Assim, este foi o procedimento adotado para a construção de uma versão de

[37] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

(VILARES; THELWALL; ALONSO, 2015) para sua utilização nos experimentos deste trabalho.

4.5 CONSIDERAÇÕES FINAIS

Este capítulo se concentrou em apresentar os detalhes referentes ao *framework* SentiBR para pré-processamento focado em mídias sociais em Português Brasileiro, com ênfase em melhorar a tarefa de Análise de Sentimento neste cenário. Primeiramente, a agenda de pesquisa adotada para o desenvolvimento do projeto foi apresentada. Em seguida, foram ilustrados com detalhes os componentes gerais da arquitetura, além de cada tarefa de pré-processamento de dados. Ressaltou-se também o impacto gerado pelas mesmas em exemplos textuais reais, a fim de se verificar como uma entrada era transformada para o módulo de classificação de sentimento.

Explicou-se o modelo de classificação escolhido, o qual foi um algoritmo SVM. Dado que esta proposta possui uma variedade de opções e parâmetros para execução, também foram ilustradas as diversas opções que um usuário pode selecionar para executar o *framework*.

Por fim, os pacotes e bibliotecas de *software*¸ bem como ferramentas utilizadas para o desenvolvimento da proposta por esse trabalho, foram explorados. O próximo capítulo irá apresentar o impacto desta arquitetura desenvolvida, quanto ao seu desempenho comparado com os *baselines* selecionados para investigação.

5 TESTES E RESULTADOS

5.1 CONSIDERAÇÕES INICIAIS

Para uma avaliação de desempenho justa e confiável da proposta presente nesta dissertação, buscou-se a aquisição de bases de dados manualmente anotadas quanto ao seu sentimento. As fontes destas bases também foram selecionadas de forma a se variar os diferentes cenários representativos de mídias sociais, como Twitter e Facebook.

Os resultados obtidos neste capítulo refletem a necessidade de se analisar cuidadosamente o comportamento das propostas experimentadas. Para isso, foram utilizadas métricas adotadas como padrão na literatura de aprendizado de máquina.

A metodologia empregada inicia pela apresentação do desempenho geral do SentiBR comparado aos demais *baselines* da literatura. Demais resultados incluem análises relacionadas a tempo de execução e desempenho individual das tarefas de pré-processamento, bem como uma discussão sobre exemplos de saída da arquitetura. Por fim, este capítulo apresenta uma discussão dos resultados e suas implicações na literatura de AS para o Português Brasileiro.

5.2 TESTES REALIZADOS

5.2.1 Bases de Dados

No processo de escolha das bases de dados para experimentação, levou-se em consideração a variedade de tipos de dados encontrados em mídias sociais. Como a revisão de literatura apresentada na Subseção 3.3.3 ilustrou, plataformas como Twitter e Facebook, além de revisões e notícias, se fazem presentes em estudos focados em AS no contexto brasileiro.

Conforme ilustrado na Subseção 2.4, a RSO Twitter se caracteriza por seu dinamismo e imensa variedade de tópicos que podem ser discutidos. A limitação quanto ao número de *caracteres* que podem ser escritos torna as instâncias desta plataforma geralmente mais curtas. Consequentemente, desafios como abreviações são comumente encontrados, sendo padrões adotados por usuários que tentam passar sua mensagem em um curto espaço.

Por outro lado, o Facebook não impõe limitação de tamanho em seus comentários escritos por usuários, o que tende a gerar instâncias com maior número de palavras ou *tokens*. Além disso, tais comentários são usualmente escritos em *fan pages*, o que pode gerar dados mais relacionados a um dado tópico referente à respectiva postagem. Porém, a variabilidade de assuntos a serem discutidos também é grande, impulsionada pela existência de *fan pages* em diversos domínios.

Tanto para o Twitter quanto para o Facebook, assume-se neste estudo que os dados advindos de tais redes são escritos de cunho pessoal, ou seja, o sentimento atrelado às instâncias destas bases é atribuído ao autor de cada comentário ou *tweet*.

Assim, as bases selecionadas para aplicação nesta dissertação estão inseridas no contexto de mídias sociais e apresentam desafios advindos de tais cenários aos modelos computacionais de AS. A Tabela 5.1 apresenta um panorama das bases adotadas para a avaliação de desempenho neste trabalho.

TABELA 5.1 PANORAMA DAS BASES DE DADOS EMPREGADAS NOS TESTES.

Autor	Tipo de Dado	Descrição	Documentos	Documentos Positivos	Documentos Negativos	Tokens	Tamanho médio de Documentos	Tamanho mínimo	Tamanho Máximo
(TELES; SANTOS; SOUZA, 2016)	Twitter	Empresas de telecomunicação	332	166	166	2116	20	2	43
(CIRQUEIRA, *et al*. 2016)	Twitter	Instituições públicas	300	150	150	2079	20	2	38
	Facebook	Instituições públicas e temas de preconceitos em RSO	250	125	125	2018	23	1	186
(NARR; HULFENHAUS; ALBAYRAK, 2012)	Twitter	Empresas diversas	1040	626	414	5066	18	2	62
(NASCIMENTO, *et al*., 2012)	Twitter	Entretenimento, Policial e Política	926	486	440	4397	21	2	46
(DOSCIATTI; FERREIRA; PARAISO, 2013)	Site de Notícias	Notícias	1499	499	1000	8990	43	14	137
(MARTINAZZO; DOSCIATTI; PARAISO, 2011)	Site de Notícias	Notícias	1002	395	607	7797	41	19	60

A primeira base vem do estudo de (TELES; SANTOS; SOUZA, 2016). Esta base foi obtida a partir de um site disponibilizado pelos autores do artigo. A mesma contém

reclamações de consumidores postadas no Twitter. Estas reclamações faziam menções a empresas de telecomunicação, tal como operadoras de telefonia móvel do Brasil.

A segunda e a terceira bases vêm do estudo de (CIRQUEIRA, *et al*. 2016). Os autores deste trabalho disponibilizaram as suas bases para este estudo[38]. As bases utilizadas são compostas de *tweets* e comentários do Facebook, relacionados às instituições: Universidade Federal do Pará (UFPA), Prefeitura de Belém, Prefeitura de Curitiba e a *fan page* Humaniza Redes. Assim, os mesmos contêm elogios e reclamações sobre o funcionamento da UFPA, serviços de prefeituras e comentários referentes a preconceito em RSO.

A quarta base adotada vem de (NARR; HULFENHAUS; ALBAYRAK, 2012). A obtenção desta base foi através de um *website* onde o autor disponibiliza um *link* para *download* da respectiva coleção. A base original é composta de *tweets* em quatro línguas diferentes, incluindo o português, a qual foi a base selecionada. Os temas são relacionados a empresas de setores diversos, tal como eletrônicos, calçados, hotelaria, automóveis e tecnologia.

A quinta coleção foi selecionada a partir do estudo de (NASCIMENTO, *et al.,* 2012). Esta base foi obtida a partir de contato realizado com os autores, via *email*. A mesma é composta de *tweets* sobre notícias que estavam em destaque no Brasil, na época em que o respetivo trabalho foi desenvolvido. Os temas centrais das notícias com as quais os *tweets* se relacionavam eram: Entretenimento, Policial e Política.

As duas últimas coleções textuais escolhidas vêm das propostas de (DOSCIATTI; FERREIRA; PARAISO, 2013) e (MARTINAZZO; DOSCIATTI; PARAISO, 2011). Estas bases foram obtidas por meio de um endereço eletrônico disponibilizado pelos autores de tais artigos. As mesmas possuem notícias advindas de um website popular no contexto jornalístico brasileiro. Apesar de que os dados estavam categorizados por emoção, foi possível realizar uma transformação para as categorias afetiva positiva e negativa, por meio de uma orientação encontrada no trabalho de (PORIA, *et al.,* 2014). Neste, os autores apresentaram a Ampulheta das Emoções, um modelo que tenta mapear diferentes emoções para uma dada polaridade, positiva ou negativa. Assim, em ambas as bases, as instâncias classificadas com as emoções de Alegria e Surpresa foram

[38] https://github.com/dougcirqueira/sentiment-analysis-social-br

consideradas positivas, e aquelas com as emoções de Desgosto, Medo, Raiva e Tristeza, receberam o rótulo negativo.

Assim, nota-se a partir da tabela anterior, que as três primeiras bases são balanceadas, enquanto que as quatro últimas têm um número diferente de instâncias positivas e negativas. O objetivo com isso era verificar o comportamento do *framework* desenvolvido em ambas as situações. A base com maior número de instâncias é a de (DOSCIATTI; FERREIRA; PARAISO, 2013), e esta também apresenta o maior número de *tokens*, ou seja, tem o maior vocabulário dentre as bases, bem como a maior média quanto ao tamanho de instâncias individuais. Porém, notou-se que a base do Facebook de (CIRQUEIRA, *et al.* 2016) foi a que apresentou maior variação no tamanho do menor para o maior documento; isto reflete o cenário usualmente encontrado no Facebook, onde os usuários têm liberdade para escrever textos maiores.

Outra observação é que as bases de (DOSCIATTI; FERREIRA; PARAISO, 2013) e (NARR; HULFENHAUS; ALBAYRAK, 2012), são originadas de notícias, o que não são dados informais de redes sociais. No caso destas, ao contrário do mencionado para as bases de RSO, que possuem sentimento atrelado ao autor de cada comentário ou *tweet*, assume-se que o sentimento associado às mesmas advém do ponto de vista do leitor de cada notícia. Dessa forma, existem exemplos que podem incitar tanto um sentimento positivo quanto negativo no leitor de cada uma das mesmas. Além disso, com estas bases é possível também avaliar a arquitetura proposta no cenário formal, sem os desafios comumente encontrados em mídias sociais.

5.2.2 Configurações dos Experimentos

Para a avaliação de desempenho do *framework* desenvolvido e demais testes realizados, adotou-se o método de validação cruzada estratificada, do inglês *k-fold cross validation* (KOHAVI, 1995) e (KATSIS, CHRISTOS D., et al.,2008). Neste contexto, foi adotado um valor *k* igual a 10 para o número de rodadas da validação.

Antes da execução da validação cruzada, foram obtidos índices únicos para todos os grupos de treinamento e teste em cada rodada, para todas as bases de dados. Assim, durante todos os testes, as instâncias eram as mesmas em cada uma das dez rodadas na validação cruzada para todos os modelos de AS. As métricas adotadas para avaliação foram acurácia, precisão, revocação e macro F1.

A partir dessas definições de configuração para experimentação e dos passos de pré-processamento a serem aplicados no SentiBR, avançou-se para os próximos testes, cujos resultados estão ilustrados na Subseção 5.3 a seguir.

5.3 RESULTADOS

Esta subseção apresenta os resultados obtidos e suas discussões divididas em subseções dedicadas. A primeira, em 5.3.1, introduz a avaliação geral da arquitetura em comparação com os *baselines* da literatura. A Subseção 5.3.2, apresenta o desempenho computacional dos modelos estudados e das tarefas de pré-processamento, que são avaliadas individualmente. Por fim, em 5.3.3 tem-se a discussão sobre exemplos de classificações equivocadas pela arquitetura e potenciais razões para este comportamento.

5.3.1 Desempenho Geral

O primeiro teste visou analisar o desempenho da arquitetura desta dissertação e dos demais estudos considerados *baselines*. Para estes resultados, a arquitetura desenvolvida foi executada com a aplicação de todas as suas tarefas de pré-processamento, e o mesmo foi realizado para os demais modelos comparativos. O objetivo desta avaliação é testar a hipótese primordial deste trabalho, que verifica se uma arquitetura expandida de pré-processamento pode melhorar resultados para AS no cenário PT_Br.

TABELA 5.2 DESEMPENHO GERAL DA ARQUITETURA DE PRÉ-PROCESSAMENTO E *BASELINES*.

Modelo	Acurácia	Precisão Positiva	Precisão Negativa	Revocação Positiva	Revocação Negativa	F1
(TELES; SANTOS; SOUZA, 2016) – Twitter – 332 Instâncias						
SentiBR	**90%**	**91,24%**	**90%**	**89,19%**	**91%**	**90%**
(TELLEZ; ERIC, *et al.*,2017a)	74%	77%	73%	70%	78%	74%
(BALAHUR; PEREA-ORTEGA, 2015)	80,09%	81%	81,35%	81%	80%	80%
(VILARES; THELWALL; ALONSO, 2015)	64%	62%	67%	72%	56%	64%

TABELA 5.2 DESEMPENHO GERAL DA ARQUITETURA DE PRÉ-PROCESSAMENTO E *BASELINES*. (CONT.)

Modelo	Acurácia	Precisão Positiva	Precisão Negativa	Revocação Positiva	Revocação Negativa	F1
(CIRQUEIRA, *et al.* 2016) - Twitter – 300 Instâncias						
SentiBR	**71%**	**72,33%**	**70,11%**	**67%**	**75%**	**70,41%**
(TELLEZ; ERIC, *et al.*,2017a)	53%	55%	52%	47%	59%	52%
(BALAHUR; PEREA-ORTEGA, 2015)	65,33%	69%	64%	57,31%	73,06%	64,27%
(VILARES; THELWALL; ALONSO, 2015)	53%	53%	53%	60%	46%	52%
(CIRQUEIRA, *et al.* 2016) - Facebook – 250 Instâncias						
SentiBR	**78%**	77,13%	**80,43%**	**81,02%**	74,10%	**77,2%**
(TELLEZ; ERIC, *et al.*,2017a)	76%	**81%**	73%	69%	**82%**	75%
(BALAHUR; PEREA-ORTEGA, 2015)	77%	76%	78,22%	79,29%	74,23%	77%
(VILARES; THELWALL; ALONSO, 2015)	67%	66%	69%	71%	63%	67%
(NARR; HULFENHAUS; ALBAYRAK, 2012) – Twitter – 1040 Instâncias						
SentiBR	**77%**	**78,36%**	**75,18%**	**86%**	**64%**	**75,07%**
(TELLEZ; ERIC, *et al.*,2017a)	68%	72%	61,30%	77,31%	54%	66%
(BALAHUR; PEREA-ORTEGA, 2015)	74%	76,14%	70,22%	83%	61%	72,07%
(VILARES; THELWALL; ALONSO, 2015)	68%	75%	59,12%	71%	**64%**	67%

TABELA 5.2 DESEMPENHO GERAL DA ARQUITETURA DE PRÉ-PROCESSAMENTO E *BASELINES*. (CONT.)

Modelo	Acurácia	Precisão Positiva	Precisão Negativa	Revocação Positiva	Revocação Negativa	F1
(NASCIMENTO, ET AL., 2012) – Twitter – 926 Instâncias						
SentiBR	**65%**	**67%**	**63,36%**	**69%**	60%	**63,07%**
(TELLEZ; ERIC, *et al*.,2017a)	59%	61,12%	56,45%	62%	56%	58%
(BALAHUR; PEREA-ORTEGA, 2015)	64%	66%	63,14%	65%	**62,27%**	62%
(VILARES; THELWALL; ALONSO, 2015)	57,20%	60,28%	55%	58,01%	56,36%	57%
(DOSCIATTI; FERREIRA; PARAISO, 2013) – Notícias – 1499 Instâncias						
SentiBR	84%	82,46%	85,11%	65,30%	**93,10%**	80%
(TELLEZ; ERIC, *et al*.,2017a)	**85%**	**83%**	**87%**	**69%**	92,4%	**81,19%**
(BALAHUR; PEREA-ORTEGA, 2015)	83%	81,10%	84,33%	63,09%	93%	79%
(VILARES; THELWALL; ALONSO, 2015)	66%	48,22%	72,26%	39,47%	79%	59,31%
(MARTINAZZO; DOSCIATTI; PARAISO, 2011) – 1002 Instâncias						
SentiBR	**68%**	**64%**	**70%**	**45%**	**82,35%**	**64%**
(TELLEZ; ERIC, *et al*.,2017a)	64%	57%	67,28%	42,04%	78%	60%
(BALAHUR; PEREA-ORTEGA, 2015)	67%	62%	69%	43,28%	82%	63%
(VILARES; THELWALL; ALONSO, 2015)	61%	51%	66%	40%	75%	57.06%

Com estes resultados é possível computar um *ranking* geral de desempenho da arquitetura desenvolvida versus as demais avaliadas. Este é baseado na métrica F1 média obtida por todos os modelos, em cada base de dados. A Tabela 5.3 abaixo ilustra este resultado, o total de acertos obtidos por cada modelo, bem como o ganho em quantidade de classificações corretas a mais por parte do SentiBR, em comparação com cada *baseline*.

TABELA 5.3 *RANKING* DE MODELOS ESTUDADOS A PARTIR DA F1 MÉDIA.

Ranking	Modelo	F1 Média Geral	Total de Acertos	Ganho Quantitativo do SentiBR
1	**SentiBR**	74%	4036	-
2	**(BALAHUR; PEREA-ORTEGA, 2015)**	71%	3921	+115
3	**(TELLEZ; ERIC, *et al*.,2017a)**	67%	3747	+289
4	**(VILARES; THELWALL; ALONSO, 2015)**	60%	3371	+665

A partir da Tabela 5.2, primeiramente nota-se que a arquitetura desta dissertação apresenta os melhores resultados para 6 das 7 bases de dados, considerando as métricas de acurácia e medida F1. Isto é refletido na Tabela 5.3, que destaca o *framework* desenvolvido como primeiro colocado, quanto à macro F1 média geral. Além disso, nota-se que o mesmo acertou 115 vezes a mais que o segundo colocado, (BALAHUR; PEREA-ORTEGA, 2015). Esta diferença em ganho quantitativo é ainda maior para (TELLEZ; ERIC, et al.,2017a) e (VILARES; THELWALL; ALONSO, 2015), para os quais o ganho foi de 289 e 665 acertos a mais, respectivamente.

A Tabela 5.2 também destaca que, para todas as coleções oriundas de mídias sociais, a arquitetura SentiBR foi superior. Além disso, a base em que o método proposto não foi superior foi a de (DOSCIATTI; FERREIRA; PARAISO, 2013), a qual é composta de notícias jornalísticas, o que não é o domínio foco das tarefas de pré-processamento analisadas nesta dissertação. Ainda sobre este resultado, quando não foi superior, a proposta ficou em segundo lugar, com F1 de 80% contra 81.19% de (TELLEZ; ERIC, *et al*.,2017a).

Ainda quanto à métrica F1, o maior desempenho da arquitetura desenvolvida foi apresentado para a base de *tweets* de (TELES; SANTOS; SOUZA, 2016), com uma F1 de 90%, contra 80% da segunda colocada, que foi (BALAHUR; PEREA-ORTEGA,

2015). Quanto ao menor desempenho, o mesmo veio da base de *tweets* de (NASCIMENTO, *et al*., 2012), onde a F1 da proposta foi 63.07%. Analisando esta base, foi possível notar que a mesma é a que mais apresenta *tokens* mal formatados. Um exemplo é o caso de URLs. Este tipo de elemento tem um formato específico, usualmente como "http://endereço.com.br". Todavia, nesta base notou-se um número elevado de URLs quebradas, como por exemplo, a presença apenas da parte "com.br". Desta forma, concluiu-se que esta era a base mais desafiadora. Mesmo assim, a arquitetura foi capaz de superar o estado da arte em tal cenário, com a maior acurácia e macro F1 em comparação com as demais.

Assim, no caso das métricas de precisão e revocação, notou-se a superioridade absoluta desta arquitetura em 4 das 7 bases, as quais são os *tweets* de (TELES; SANTOS; SOUZA, 2016), (CIRQUEIRA, *et al*. 2016), (NARR; HULFENHAUS; ALBAYRAK, 2012) e as notícias de (MARTINAZZO; DOSCIATTI; PARAISO, 2011). Ou seja, nestes casos o *framework* foi capaz de tomar decisões corretas por mais vezes, quando da análise individual de sentimento nas classes positiva e negativa.

Quanto à precisão do modelo, observou-se nas bases oriundas de mídias sociais que a arquitetura proposta sempre apresenta maior precisão para a classe positiva. Tal padrão não é encontrado para a métrica de revocação, que por vezes é maior para a classe positiva, outrora negativa. Já no caso das duas últimas bases, encontrou-se um padrão de maior precisão e revocação para a classe negativa.

Neste contexto, vale ressaltar que estas duas últimas bases são compostas de notícias, o que são usualmente escritas de forma imparcial. Entretanto, após análise destas bases de dados, notou-se que as notícias com emoções negativas empregavam mais palavras negativas, enquanto que a positividade não era expressa com essa característica estatística nas notícias positivas. Assim, acredita-se que esta pode ser uma razão para tal comportamento do modelo. Para este estudo específico, buscou-se por palavras positivas e negativas em todas as notícias, a partir da lista que a arquitetura utiliza na função de captura de palavras afetivas, mencionada na Subseção 4.3.

Analisando a revocação, observou-se que a arquitetura apresentou o maior número de verdadeiros positivos e verdadeiros negativos, em comparação com todos os demais métodos, totalizando 1774 e 2312 acertos, respectivamente. Ou seja, apesar de não ter superado os concorrentes em todas as 7 bases, esta arquitetura foi capaz de acertar mais

quando somadas todas as instâncias presentes na experimentação. Estes resultados também representam um ganho médio de 3% quanto a métrica de revocação para o SentiBR, em comparação com a revocação do estado da arte.

Assim, dados os resultados supracitados, pode-se confirmar uma tendência na melhoria dos resultados de AS para o PT_Br, quando se aplica um conjunto expandido de tarefas de pré-processamento.

5.3.2 Avaliação de Desempenho Computacional

Essa subseção avalia o desempenho computacional em dois momentos. Primeiro, quanto aos tempos de pré-processamento do *framework* da dissertação e dos demais *baselines* analisados. Segundo, a mesma apresenta o panorama do desempenho individual de cada tarefa de pré-processamento do *framework* desenvolvido. Assim, a Tabela 5.4 apresenta os tempos de pré-processamento em minutos para cada modelo, os quais estão organizados em ordem de desempenho, quanto a sua macro F1 obtida. O objetivo é ilustrar o *trade-off* entre tempo e desempenho, para cada método e base de dados.

TABELA 5.4 DESEMPENHO COMPUTACIONAL POR MODELO E BASE.

Modelo	Tempo em Minutos	Métrica F1
(TELES; SANTOS; SOUZA, 2016) – Twitter – 332 Instâncias		
SentiBR	27,24	90%
(BALAHUR; PEREA-ORTEGA, 2015)	3,72	80%
(TELLEZ; ERIC, *et al*.,2017a)	117,48	74%
(VILARES; THELWALL; ALONSO, 2015)	0,0157	64%
(CIRQUEIRA, *et al.* 2016) - Twitter – 300 Instâncias		
SentiBR	30,13	70,41%
(BALAHUR; PEREA-ORTEGA, 2015)	3,29	64,27%
(TELLEZ; ERIC, *et al*.,2017a)	126,45	52,00%
(VILARES; THELWALL; ALONSO, 2015)	0,0131	52,00%
(CIRQUEIRA, *et al.* 2016) - Facebook – 250 Instâncias		
SentiBR	27,49	77,02%
(BALAHUR; PEREA-ORTEGA, 2015)	2,72	77,00%
(TELLEZ; ERIC, *et al*.,2017a)	98,27	75,00%
(VILARES; THELWALL; ALONSO, 2015)	0,0122	67,00%
(NARR; HULFENHAUS; ALBAYRAK, 2012) – Twitter – 1040 Instâncias		
SentiBR	64,83	75,07%
(BALAHUR; PEREA-ORTEGA, 2015)	11,11	72,07%
(VILARES; THELWALL; ALONSO, 2015)	1,36	67%
(TELLEZ; ERIC, *et al*.,2017a)	107	66%

TABELA 5.4 DESEMPENHO COMPUTACIONAL POR MODELO E BASE. (CONT.)

Modelo	Tempo em Minutos	Métrica F1
(NASCIMENTO, *et al*., 2012) – Twitter – 926 Instâncias		
SentiBR	71,06	63,07%
(BALAHUR; PEREA-ORTEGA, 2015)	10,42	62%
(TELLEZ; ERIC, *et al*.,2017a)	99	58%
(VILARES; THELWALL; ALONSO, 2015)	1,32	57%
(DOSCIATTI; FERREIRA; PARAISO, 2013) – Notícias – 1499 Instâncias		
(TELLEZ; ERIC, *et al*.,2017a)	117	81,19%
SentiBR	162,59	80%
(BALAHUR; PEREA-ORTEGA, 2015)	18	79%
(VILARES; THELWALL; ALONSO, 2015)	1,69	59,31%
(MARTINAZZO; DOSCIATTI; PARAISO, 2011) – Notícias – 1002 Instâncias		
SentiBR	108,28	64%
(BALAHUR; PEREA-ORTEGA, 2015)	12	63%
(TELLEZ; ERIC, *et al*.,2017a)	102	60%
(VILARES; THELWALL; ALONSO, 2015)	1,63	57,06%

A Tabela 5.5 sumariza os resultados apresentados previamente para os quatro modelos em comparação, fornecendo o tempo médio de execução e macro F1 média, sobre todas as bases de dados. A disposição nesta tabela está por ordem de desempenho.

TABELA 5.5 MÉDIA DE TEMPO DE EXECUÇÃO E DESEMPENHO POR BASE.

Modelo	Minutos	Métrica F1
SentiBR	70,23142857	74%
(BALAHUR; PEREA-ORTEGA, 2015)	8,751428571	71%
(TELLEZ; ERIC, *et al*.,2017a)	109,6	67%
(VILARES; THELWALL; ALONSO, 2015)	0,863	60%

O gráfico de bolhas a seguir fornece uma nova forma de observação da Tabela 5.5. Assim, é possível enriquecer a visualização do *trade-off* tempo versus desempenho.

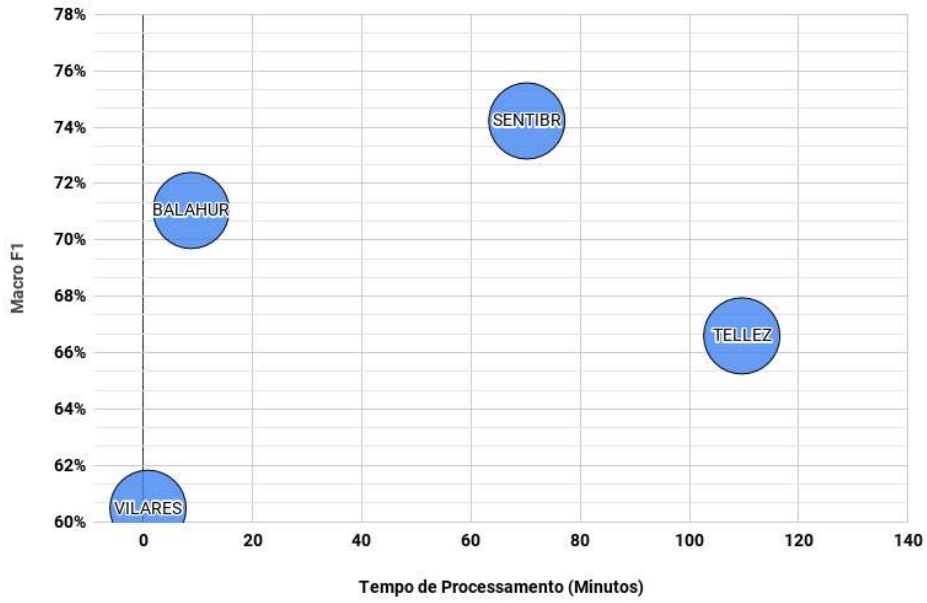

Figura 5.1 Tempo de Pré-processamento por desempenho das propostas.

A partir dos resultados obtidos, é possível observar que nem sempre a proposta a executar mais rapidamente será a mais eficiente em desempenho. Por exemplo, a Tabela 6.5 ilustra que (VILARES; THELWALL; ALONSO, 2015) executou seu pré-processamento completo para todas as bases com um tempo médio de 0.86 minutos, mas foi o pior no *ranking* de desempenho geral. Por outro lado, a proposta com maior tempo de execução também não foi a melhor, que é o caso da terceira colocada no *ranking* de desempenho, (TELLEZ; ERIC, *et al.*,2017a), com tempo médio de 109.6 minutos para um pré-processamento completo de todas as bases.

Quanto à arquitetura desenvolvida, a mesma ficou em primeiro por desempenho, porém em terceiro quanto ao tempo de execução. Dessa forma, cabe ao usuário da aplicação avaliar o *trade-off* entre o melhor desempenho e o tempo necessário de pré-processamento.

A fim de investigar onde se encontrava esta alta impedância quanto ao tempo de execução na arquitetura desta dissertação, a próxima análise concentrou-se em averiguar o desempenho computacional individual de pré-processamento por cada tarefa de tratamento de dados.

Neste segundo momento, são apontados o tempo total, as médias de tempo e desempenho por cada tarefa de pré-processamento empregada na arquitetura.

Para a apresentação destes resultados, coletou-se o tempo total de execução por cada tarefa de pré-processamento, em cada base de dados. Este tempo total a partir de cada base foi então somado, e o mesmo é ilustrado na coluna "Tempo Total", na Tabela 5.6. Em seguida, o tempo total da coluna anterior foi dividido pelo total de instâncias de todas as sete bases de dados. Estes resultados da média de tempo de execução por todas as coleções estão presentes na coluna "Tempo Médio". Devido a esta média, optou-se pela ilustração do tempo em segundos, dado que a conversão para minutos tornaria difícil a visualização de tais resultados, por se tornarem valores muito pequenos.

Para compor a última coluna desta tabela, a arquitetura desenvolvida foi testada utilizando-se apenas uma tarefa de pré-processamento por vez, para todas as bases de dados. A cada execução, a macro F1 obtida era salva. Assim, foi calculada a média geral dessa métrica, para cada tarefa em todas as bases. Estes resultados então compõem a coluna Macro F1 Média na tabela.

A Tabela 5.6 é apresentada em ordem decrescente de macro F1 média obtida de cada tarefa de pré-processamento.

TABELA 5.6 *RANKING* DE TAREFAS DE PRÉ-PROCESSAMENTO POR SEU DESEMPENHO.

Tarefa	Tempo Total	Tempo Médio	Macro F1 média
Stemming	6,05	0,001130	**0,73090**
Minúsculas	0,02	0,000003	0,72758
Correção Ortográfica Fonética	**29084,18**	**5,437312**	0,72616
Tokenização	2,75	0,000514	0,72573
Palavras Afetivas	207,95	0,038877	0,72343
Xingamento	6,84	0,001279	0,71898
Stopwords	0,44	0,000083	0,71760
Repetição de Pontuação	0,37	0,000069	0,71633
Acentos	0,11	0,000021	0,71612
Números	1,35	0,000252	0,71401
Emoticons/Emojis	14,84	0,002774	0,71376
Cumprimentos	0,12	0,000023	0,71247
Amplificadores/Detratores	0,96	0,000180	0,71222

TABELA 5.6 *RANKING* DE TAREFAS DE PRÉ-PROCESSAMENTO POR SEU DESEMPENHO. (CONT.)

Tarefa	Tempo Total	Tempo Médio	Macro F1 média
Negação	32,55	0,006085	0,71214
Abreviações	1,40	0,000261	0,71179
Gírias	1,68	0,000314	0,71154
Advérbios de Dúvida	0,17	0,000031	0,71104
Maiúsculas	0,23	0,000043	0,71057
Risadas	0,99	0,000186	0,70879
Repetições de Letras	0,85	0,000159	0,70748
***Hashtags*/Menções/URLs**	0,42	0,000078	0,69523
REN	68,97	0,012895	-
PoS	65,16	0,012181	-

Nesta tabela, são destacados em negrito o maior tempo total, tempo médio e macro F1. Analisando as primeiras colunas, foi evidenciado que a tarefa com mais tempo de processamento era a de Correção Ortográfica Fonética, com um total de 29084,18 segundos e tempo médio de aproximadamente 5 segundos. A mesma foi seguida por Palavras Afetivas, Reconhecimento de Entidades Nomeadas, *Part-of-Speech tagging*, Negação, *Emoticons & Emojis*. As demais tarefas apresentaram tempo de execução total abaixo de 7 segundos.

Quanto ao desempenho individual de cada tarefa de tratamento, notou-se que o *Stemming* forneceu maior contribuição, com uma F1 média de 73.09%. Acredita-se que a razão para este resultado é atribuída pela tarefa gerar uma normalização textual que atinge toda a base. Isto é diferente, por exemplo, da tarefa de Palavras Afetivas, que apesar de estar entre as 5 principais por performance, realiza a normalização somente nas palavras encontradas em sua lista.

Ainda no quesito desempenho individual, o *Stemming* foi seguido pelas tarefas de Minúsculas, Correção Ortográfica Fonética e Tokenização, e a já mencionada Palavras Afetivas, todas com F1 média acima de 72%. Quanto às demais tarefas, nota-se que 5 das mesmas que não estão presentes em trabalhos relacionados, apresentaram macro F1 acima de 71%, a saber: Correção Ortográfica Fonética, Xingamentos, Cumprimentos, Advérbios de Dúvida, e Maiúsculas. Isto revela a necessidade e importância de tais tarefas, quando se trabalha com dados de mídias sociais para AS no contexto PT_Br.

É interessante notar que a tarefa posicionada por último foi relacionada ao tratamento de *hashtags*, menções e URLs, pois estas são frequentemente encontradas em

trabalhos presentes no contexto brasileiro. Acredita-se que a razão para este resultado se deve ao fato de as *hashtags* presentes nos dados nem sempre apresentarem alguma informação de conotação afetiva, algo que já é esperado de menções e URLs. Todavia, quando aplicada com as demais tarefas em conjunto, esta também se faz necessária para se alcançar o desempenho mais elevado em 5 das 7 bases de dados, como será ilustrado nos próximos resultados.

Outro ponto a se observar, é que dentre as tarefas com maior contribuição em macro F1, duas estão entre as que mais consomem tempo de processamento na arquitetura, a saber: Correção Ortográfica Fonética e Palavras Afetivas. Assim, apesar de um tempo maior ser necessário para processá-las, cabe novamente ao usuário da aplicação avaliar a necessidade das mesmas e este *trade-off* tempo e desempenho individual das tarefas.

Assim, este *ranking* também pode contribuir para revelar quais tarefas devem ser consideradas com maior peso para realizar a classificação de sentimento em dados oriundos de RSO e demais plataformas de mídia social brasileiras.

5.3.3 Exemplos de Classificação Errônea

Com o objetivo de se investigar potenciais razões para os erros de classificação por parte da arquitetura proposta, a Tabela 5.7 apresenta instâncias que receberam a classe errada quanto ao seu sentimento. A mesma ilustra também como o texto foi transformado após as tarefas de pré-processamento.

TABELA 5.7 EXEMPLOS DE INSTÂNCIAS CLASSIFICADAS ERRONEAMENTE.

Entrada	Saída	Classe	Predição
Vendas 'mornas' do Wii derrubam lucro da Nintendo: Queda no trimestre foi de 61% em comparação com 2008. Consoles da empresa japonesa ainda lideram o mercado.	vend ' morn ' wii derrub positive intend qued trimestr number comparaca number positive negative japones aind lid merc	Positivo	Negativo
O estagiário da @Vivoemrede tropeçou no fio do 3G.	estagiari mention tropec fio 3g	Negativo	Positivo
To na loja da @vivo comprando um 'BoxVivo'. Negócio tá vendendo como água	loj mention compr ' boxviv ' negoci vend negative	Positivo	Negativo
Perfilzinho de partido	perfilzinh part	Negativo	Positivo

A primeira é um exemplo de notícia positiva, que foi considerada negativa. Nota-se que a positividade neste texto vem no fim, quando o mesmo menciona que o lucro da Nintendo se manteve em alta. Todavia, esta informação é inserida diretamente, sem a utilização de uma conjunção adversativa. Isto pode ter prejudicado a tarefa de Negação, que poderia capturar esta inversão de polaridade na parte final do texto.

Os segundo e terceiro exemplos são *tweets*, e em ambos os casos, é notório que o sentimento negativo e positivo, respectivamente, advém de expressões populares, como "tropeçar" e "vender como água". Estas são expressões não capturadas pelos recursos existentes da arquitetura. Além disso, notou-se que a palavra "água" foi substituída pela tarefa de Palavras Afetivas, por um identificador "negative". Isto aparenta não ser um identificador preciso para esta palavra. Assim, ressalta-se aqui que o *framework* permite a atualização de seus recursos por parte de seus usuários, para sempre mantê-la habilitada a capturar novas gírias, idiomas e o sentimento correto atrelado a diferentes palavras no português.

O último exemplo vem de um comentário do Facebook. Neste caso, a negatividade está presente no uso do diminutivo para o termo "perfil". Este é um potencial sinal indicador de polaridade, a ser considerado como modo de ponderação para futuras aplicações do *framework* desenvolvido.

6 CONCLUSÃO

É inegável a importância da Análise de Sentimento para estudos envolvendo UGC em mídias sociais. A relevância de mercado dessa técnica tem crescido, impulsionada pela onipresença de Redes Sociais Online no cotidiano de consumidores. O Brasil é um cenário exemplo, onde se observa este fenômeno diariamente, dado a grande quantidade de acessos e dados postados por brasileiros, seja em perfis pessoais ou de suas marcas e empresas favoritas.

No entanto, percebeu-se na literatura de Análise de Sentimento no contexto brasileiro, a ausência de uma padronização em uma etapa essencial desta técnica: o pré-processamento. Este passo visa realizar o tratamento dos dados antes da aplicação das etapas de classificação automática de sentimento. A mesma pode ter impacto direto na precisão e confiabilidade da AS em relação ao PT_Br, principalmente de dados provenientes de RSO.

Além disso, entre as constatações, percebeu-se que a maioria das propostas na literatura implementam tarefas menos complexas e com ênfase na remoção de ruídos, que é o caso da eliminação de *stopwords* e tokenização. Também foi observado que existem várias metodologias para implementar a mesma tarefa de pré-processamento. No entanto, poucos autores brasileiros exploram essa diversidade de abordagens.

A variação relativa quanto ao número de tarefas realizadas em cada proposta também foi notada. Em geral, cada estudo apresenta tarefas específicas de pré-processamento, com foco em sua aplicação. Este fato reforçou a conclusão de que não existe um *framework* uniforme para o pré-processamento dos dados das mídias sociais, com ênfase em Análise de Sentimento na língua portuguesa brasileira.

Este trabalho visou responder a hipótese de que: uma arquitetura expandida de pré-processamento para conteúdo gerado por consumidores usuários em Redes Sociais Online, pode melhorar o desempenho de tarefas de Mineração de Texto, tal como Análise de Sentimento neste domínio. Para tal, foi desenvolvida uma arquitetura de pré-processamento cobrindo um maior número de tarefas de tratamento de dados, e que não havia ainda sido proposta no estado da arte, denominada SentiBR. Sete bases de dados de mídias sociais e notícias foram coletadas, a fim de se realizar uma avaliação de desempenho da arquitetura desenvolvida.

Para a comparação com a literatura existente, foram selecionadas três metodologias do estado da arte, consideradas bem estabelecidas no contexto de mídias sociais para linguagens românticas, assim como o PT_Br.

Um primeiro experimento analisou a performance geral da arquitetura, com a aplicação de todas as 23 tarefas de pré-processamento. O mesmo procedimento foi repetido para os demais modelos comparativos. Notou-se então a tendência de que, com um maior número de tarefas de tratamento, a cobrir uma gama maior de desafios da linguagem em mídias sociais, pode-se otimizar o desempenho de AS neste contexto. A prova desta conclusão foi a performance vencedora da arquitetura desta dissertação, em 6 das 7 bases adotadas para a experimentação. Notou-se também que a revocação média desta arquitetura superou o estado da arte em 3%, ou seja, o SentiBR consegue detectar um maior número de verdadeiros positivos. Todavia, um fator negativo para esta implementação foi o maior tempo de execução, advindo principalmente da tarefa de Correção Ortográfica Fonética.

Os demais experimentos ilustraram as tarefas com maior contribuição individual para o desempenho da arquitetura, as quais foram *Stemming*, conversão para minúsculas, Correção Ortográfica, tokenização, e detecção de palavras afetivas. Assim, estas podem ser potenciais indicações de tarefas que devem sempre estar inseridas em um *pipeline* de Análise de Sentimento para mídias sociais.

Finalmente, com a análise de classificação errônea para bases diversificadas, notou-se a necessidade de se melhorar o desempenho do *framework* para lidar com demais expressões populares e gírias, além de contextualização. Desta forma, conclui-se que existe a necessidade de se expandir as tarefas de pré-processamento na aplicação de Análise de Sentimento para dados de mídias sociais e Redes Sociais Online. Este estudo provê um *framework* completo para auxílio nesta tarefa, a fim de impulsionar as investigações na temática para o contexto do Português Brasileiro.

Este trabalho foi desenvolvido no contexto do grupo de pesquisa em Social CRM do Laboratório de Inteligência Computacional e Pesquisa Operacional (LINC-UFPA), o qual é focado em soluções para interação de relacionamento com clientes em mídias sociais. Esta dissertação também é parte de um projeto de Mineração de Texto para Português Brasileiro, o qual também se constitui como uma tese de doutorado da pesquisadora Márcia Fontes Pinheiro.

6.1 CONTRIBUIÇÕES

Dentre as contribuições desta proposta de dissertação, tem-se:

- Revisão da literatura sobre pré-processamento para Análise de Sentimento em Português Brasileiro;
- Metodologia para tratamento de dados em mídias sociais do Brasil, tendo como objetivo propiciar a padronização do processo de pré-processamento para Análise de Sentimento; nesse contexto, possibilitando a aceleração das pesquisas relacionadas e comparações na literatura;
- Disponibilização de uma arquitetura de pré-processamento para a execução de testes por parte da comunidade acadêmica e industrial em Análise de Sentimento, através de um link para seu repositório *git*[39].

6.2 PUBLICAÇÕES GERADAS

- Cirqueira, D., Pinheiro, M., Braga, T., Jacob Jr, A., Reinhold, O., Alt, R., Santana, Á. Improving relationship management in universities with sentiment analysis and topic modeling of social media channels: learnings from UFPA. In Proceedings of the International Conference on Web Intelligence (pp. 998-1005). ACM, 2017.
- Cirqueira, D., Vinícius, L., Pinheiro, M., Junior, A. J., Lobato, F., Santana, Á. Opinion Label: A Gamified Crowdsourcing System for Sentiment Analysis Annotation. In Proceedings of the XVI Workshop in Tools and Applications (WFA) in the XXIII Brazilian Symposium on Multimedia and Web Systems, 2017.
- Almeida, G. R., Cirqueira, D. R., Lobato, F. M. Improving Social CRM through eletronic word-of-mouth: a case study of ReclameAqui. In Proceedings of the XIV Workshop on Scientific Initiation Works (WTIC) in the XXIII Brazilian Symposium on Multimedia and Web Systems. 2017.

[39] https://gitlab.com/social-crm-linc/sentibr.git

6.3 TRABALHOS FUTUROS

Como sugestão de trabalhos futuros, tem-se:

- Emprego de metaheurística para a busca de parâmetros otimizados para as tarefas de pré-processamento;
- Aplicar demais tarefas de pré-processamento não testadas nessa versão, tal como o impacto do tempo verbal e diminutivos na polaridade final;
- Avaliação de desempenho para a representação de *Word Embeddings* (MIKOLOV, *et al*., 2013a) e (MIKOLOV, *et al.,* 2013b), atrelada à arquitetura de pré-processamento.

REFERÊNCIAS

ADAMOPOULOS, P.; GHOSE, A.; TODRI, V. The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms. Information Systems Research, 2018.

APTÉ, C., & WEISS, S. (1997). Data mining with decision trees and decision rules. Future generation computer systems, 13(2-3), 197-210.

ARCHAK, N., GHOSE, A., & IPEIROTIS, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. Management science, 57(8), 1485-1509.

ARUN, K., & NAYAGAM, M. G. (2014). Building Applications with Social Networking API's. International Journal of Advanced Networking and Applications, 5(5), 2070.

AVANÇO, L. V. (2015). Sobre normalizaçao e classificaçao de polaridade de textos opinativos na web (Doctoral dissertation, Universidade de São Paulo).

AVANÇO, L. V., BRUM, H. B., & NUNES, M. G. (2016). Improving opinion classifiers by combining different methods and resources. XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), 25-36

B. KITCHENHAM AND S. CHARTERS, "Guidelines for performing Systematic Literature Reviews in Software Engineering," Tech. Rep. EBSE 2007-001, 2007.

BACCIANELLA, S., ESULI, A. AND SEBASTIANI, F., MAY. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204), 2010.

BADAL, V.; KUNDROTAS, P.; VAKSER, I. Natural language processing in text mining for structural modeling of protein complexes. BMC bioinformatics, v. 19, n. 1, p. 84, 2018.

BAHRAINIAN, S. A., & DENGEL, A. (2013, November). Sentiment analysis using sentiment features. In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03 (pp. 26-29). IEEE Computer Society.

BALAGE FILHO, P.P., PARDO, T.A. AND ALUISIO, S.M. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL) (pp. 215-219) , 2013.

BALAHUR, A., & PEREA-ORTEGA, J. M. (2015). Sentiment analysis system adaptation for multilingual processing: The case of tweets. Information Processing & Management, 51(4), 547-556.

BEJAN, C. A., ANGIOLILLO, J., CONWAY, D., NASH, R., SHIREY-RICE, J., LIPWORTH-ELLIOT, L., ... & JOHNSON, K. B. Large-Scale Text Mining of Social

Determinants from Electronic Health Records: Case Studies of Homelessness and Adverse Childhood Experiences. In AMIA. 2017.

BEREZINA, K., BILGIHAN, A., COBANOGLU, C., & OKUMUS, F. Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. Journal of Hospitality Marketing & Management, 25(1), 1-24. 2016.

BOX, G.E. AND TIAO, G.C. Bayesian inference in statistical analysis(Vol. 40). John Wiley & Sons, 2011.

BREIMAN, L. (2001). Random forests. Machine learning, 45(1), 5-32.

BREIMAN, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

BUETTNER, R. Predicting user behavior in electronic markets based on personality-mining in large online social networks. Electronic Markets, v. 27, n. 3, p. 247-265, 2017.

CALLOU, D.; LEITE, Y.; MORAES, J. A. Variação e diferenciação dialetal: a pronúncia do /r/ no português do Brasil. In: KOCH, I. G. V. (Org.). Gramática do português falado. V. 6. Campinas: Editora da Unicamp, Fapesp, p. 465-493, 1996.

CALORE, M. “How Foursquare Is Forcing Social Networks to Check In or Check Out”. Disponível em: <http://www.wired.com/2013/03/location-apps-social-media/>. Acesso em: 3 mai. 2016

CAMBRIA, E., SPEER, R., HAVASI, C. AND HUSSAIN, A. SenticNet: A Publicly Available Semantic Resource for Opinion Mining. InAAAI fall symposium: commonsense knowledge (Vol. 10, No. 0), 2010.

CARVALHO, A., FACELI, K., LORENA, A., & GAMA, J. Inteligência Artificial–uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC. 2011.

CEJUELA, J. M., BOJCHEVSKI, A., UHLIG, C., BEKMUKHAMETOV, R., KUMAR KARN, S., MAHMUTI, S., BAGHUDANA, A., DUBEY A, SATAGOPAM, V. & ROST, B. Nala: text mining natural language mutation mentions. Bioinformatics, 33(12), 1852-1858. 2017.

CERON, ANDREA, et al. "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France." New Media & Society 16.2 (2014): 340-358.

CHAO, L.,“Brazil: The Social Media Capital of the Universe”. Disponível em: <http://www.wsj.com/articles/SB10001424127887323301104578257950857891898>. Acesso em: 05 mai. 2016

CHEN, W.; BARBOUR, R. Life priorities in the HIV-positive Asians: a text-mining analysis in young vs. old generation. AIDS care, v. 29, n. 4, p. 507-510, 2017.

CHOWDHURY, G.G., Natural language processing. Annual review of information science and technology, 37(1), pp.51-89, 2003.

CHURCH, K.W., HANKS, P.: Word Association Norms, Mutual Information and Lexicography. In: Proceedings of the 27th Annual Conference of the Association of Computational Linguistics, (1989) 76-83

CIRQUEIRA, DOUGLAS, et al. "Performance evaluation of sentiment analysis methods for Brazilian Portuguese." International Conference on Business Information Systems. Springer, Cham, 2016.

CLARK, E., & ARAKI, K. (2011). Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. Procedia-Social and Behavioral Sciences, 27, 2-11.

COOK, N. (2017). Enterprise 2.0: How social software will change the future of work. Routledge

COOLEY, R., MOBASHER, B., & SRIVASTAVA, J. (1997, November). Web mining: Information and pattern discovery on the world wide web. In Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on (pp. 558-567). IEEE.

CORRÊA, EDILSON ANSELMO, et al. "PELESent: Cross-domain polarity classification using distant supervision." Intelligent Systems (BRACIS), 2017 Brazilian Conference on. IEEE, 2017.

CORTES, C. AND VAPNIK, V., Support vector machine. Machine learning,20(3), pp.273-297, 1995.

DAI, H. J., WEI, C. H., KAO, H. Y., LIU, R. L., TSAI, R. T. H., & LU, Z. Text mining for translational bioinformatics. BioMed research international, 2015.

DAVIS; GOADRICH, 2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240). ACM.

DE ARAUJO, GABRIELA DENISE, et al. "Sentiment Analysis of Twitter's Health Messages in Brazilian Portuguese." Journal of Health Informatics 10.1 (2018).]

DEGENNE, A. AND FORSÉ, M. Introducing social networks. Sage, 1999.

DIAS, C. M. T., Aplicação do Questionário Piloto de Base Semântico-lexical do Estado do Pará/1997, Trabalho de Conclusão de Curso, Universidade Federal do Pará, Instituto de Letras e Artes, 2001

DIETTERICHL, T. G. (2002). Ensemble learning.

DOSCIATTI, M. M., FERREIRA, L. P. C., & PARAISO, E. C. (2013). Identificando emoções em textos em português do Brasil usando máquina de vetores de suporte em solução multiclasse. ENIAC-Encontro Nacional de Inteligência Artificial e Computacional. Fortaleza, Brasil.

DURAN, M. S., AVANÇO, L. V., & NUNES, M. D. G. V. (2015). A normalizer for UGC in brazilian portuguese. In Workshop on Noisy User-generated Text. Association for Computational Linguistics-ACL.

ELAGAMY, M.; STANIER, C.; SHARP, B. Text Mining Approach to Analyse Stock Market Movement. In: International Conference on Advanced Machine Learning Technologies and Applications. Springer, Cham, 2018. p. 661-670.

ELLISON, N.B. Social network sites: Definition, history, and scholarship.Journal of Computer-Mediated Communication, 13(1), pp.210-230, 2007.

FACEBOOK. "Company Info". Disponível em: <http://newsroom.fb.com/company-info/>. Acessado em: 11 mar. 2016

FACEBOOK FOR DEVELOPERS. "A Graph API". Disponível em: <https://developers.facebook.com/docs/graph-api>. Acessado em: 15 fev. 2016

FACEBOOK. "Investor Relations". Disponível em: <http://investor.fb.com/index.cfm>. Acessado em: 23 abr. 2016

FAN, W., & GORDON, M. D. (2014). The power of social media analytics. Communications of the ACM, 57(6), 74-81

FAYYAD, U., PIATETSKY-SHAPIRO, G., & SMYTH, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

FELDMAN, R. Techniques and applications for sentiment analysis. Communications of the ACM, v. 56, n. 4, p. 82-89, 2013.

FELDMAN, R.; DAGAN, I. Knowledge Discovery in Textual Databases (KDT). In: KDD. 1995. p. 112-117.

FELDMAN, R.; SANGER, J. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press, 2007.

FERSINI, E., MESSINA, E., & POZZI, F. A. (2016). Expressive signals in social media languages to improve polarity detection. Information Processing & Management, 52(1), 20-35.

FOLTRAN, M. J., & NOBREGA, V. A. (2016). Intensifier adjectives in Brazilian Portuguese: properties, distribution and morphological reflexes. Alfa: Revista de Linguística (São José do Rio Preto), 60(2), 319-340.

GALLI, F.C.S. Linguagem da Internet: um meio de comunicação global. Hipertexto e gêneros digitais: novas formas de construção de sentido. Rio de Janeiro: Lucerna, pp.120-134, 2004.

GARDNER, M. W., & DORLING, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment, 32(14-15), 2627-2636.

GARRIGOS, et al. Social networks and Web 3.0: their impact on the management and marketing of organizations. Management Decision, 50(10), 1880-1890. 2012

GHIASSI, M., SKINNER, J., & ZIMBRA, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with applications, 40(16), 6266-6282.

GIATSOGLOU, MARIA, et al. "Sentiment analysis leveraging emotions and word embeddings." Expert Systems with Applications 69 (2017): 214-224.

GILBERT, C.H.E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, 2014.

GONÇALVES, C. A. (2002). Morfopragmática da intensificação sufixal em português. Revista de Letras, 1(24).

GONZALEZ, Z. M. G. (2007). Linguística de corpus na análise do internetês.

GRANDIN, P., & ADAN, J. M. (2016). Piegas: A systems for sentiment analysis of tweets in portuguese. IEEE Latin America Transactions, 14(7), 3467-3473.

GRISHMAN, R. (1984). Natural language processing. Journal of the American Society for Information Science, 35(5), 291-296.

HAGAN, M.T., DEMUTH, H.B., BEALE, M.H. AND DE JESÚS, O., Neural network design (Vol. 20). Boston: PWS publishing company, 1996.

HANNAK, ANIKO, et al. "Tweetin'in the Rain: Exploring Societal-Scale Effects of Weather on Mood." ICWSM. 2012.

HASTIE, T., ROSSET, S., ZHU, J., & ZOU, H. (2009). Multi-class adaboost. Statistics and its Interface, 2(3), 349-360.

HATZIVASSILOGLOU, V. AND MCKEOWN, K.R., July. Predicting the semantic orientation of adjectives. In Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European chapter of the association for computational linguistics (pp. 174-181). Association for Computational Linguistics, 1997.

HEARST, MARTI A., et al. "Support vector machines." Intelligent Systems and their Applications, IEEE 13.4 (1998): 18-28.

HOSMER JR, D. W., LEMESHOW, S., & STURDIVANT, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

HU, M. AND LIU, B. Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760), 2004.

HU, M. AND LIU, B., August. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM, 2004.

HUGO GONÇALO OLIVEIRA AND PAULO GOMES. ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. In Language Resources and Evaluation 48(2):373-393. Springer, 2014.

HUTTO, C. J., & GILBERT, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. Pesquisa Nacional por Amostra de Domícilios contínua – PNAD Contínua: 2016 Acesso à Internet e à televisão e posse de telefone móvel em celular para uso pessoal, pág. 43, 2018.

INTERNET LIVE STATS. “Twitter Usage Statistics”. Disponível em: <http://www.internetlivestats.com/twitter-statistics/>. Acessado em: 29 julho. 2018

JAIN, A.K., MURTY, M.N. AND FLYNN, P.J., Data clustering: a review.ACM computing surveys (CSUR), 31(3), pp.264-323, 1999.

JONES, K.S., Natural language processing: a historical review. InCurrent issues in computational linguistics: in honour of Don Walker (pp. 3-16). Springer Netherlands, 1994.

KANNAN, S. AND GURUSAMY, V. Preprocessing Techniques for Text Mining, 2014.

KATSIS, CHRISTOS D., et al. "Toward emotion recognition in car-racing drivers: A biosignal processing approach." IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 38.3 (2008): 502-512.

KHAN, F. H., BASHIR, S., & QAMAR, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems, 57, 245-257.

KIM, Y. B., LEE, J., PARK, N., CHOO, J., KIM, J. H., & KIM, C. H. When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. PloS one, 12(5), e0177630. 2017.

KOHAVI, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145.

KOMESU, F. A modalidade escrita nas páginas eletrônicas pessoais da internet: o uso de emoticons e de “risadinhas”. Sínteses, 7, pp.167-180, 2002.

KOMESU, F. AND TENANI, L. Considerações sobre o conceito de" internetês" nos estudos da linguagem. Linguagem em (Dis) curso, 9(3), p.8, 2009.

KOMESU, F. Internetês para interneteiros:(velhas) questões sobre escrita. Estudos Linguísticos, pp.1000-1007, 2007.

KOTSIANTIS, S. B., ZAHARAKIS, I., & PINTELAS, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24.

KIBRIYA, ASHRAF M., et al. "Multinomial naive bayes for text categorization revisited." Australasian Joint Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2004.

KRUG, S. “Reactions Now Available Globally”. Disponível em: <http://newsroom.fb.com/news/2016/02/reactions-now-available-globally/>. Acessado em: 10 mar. 2016

KUSHIMA, M., ARAKI, K., YAMAZAKI, T., ARAKI, S., OGAWA, T., & SONEHARA, N. Text Data Mining of Care Life Log by the Level of Care Required Using KeyGraph. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1). 2017.

LANDAUER, T.K., Dumais, S.T.: A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104 (1997) 211-240.

LEGGAT, H., “What is Sentiment Analysis”. Disponível em: <http://www.bizreport.com/2011/05/ibm-social-media-marketing-expections-come-back-down-to-eart.html>. Acesso em: 02 mai. 2016

LEVALLOIS, C. Umigon: sentiment analysis for tweets based on lexicons and heuristics. In Proceedings of the International Workshop on Semantic Evaluation, SemEval (Vol. 13), 2013.

LI, Y. Text Mining and Financial News: Could News Sentiment affect Market Behavior?. 2017.

LING, G., YUAN, S., NI, G., YANTING, L., & QIAN, S. Clustering analysis vulnerability information based on text mining. Journal of Southeast University: Natural Science Edition, 45(5), 845-850. 2015.

LIU, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.),

LIU, B., & ZHANG, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer, Boston, MA.

LOBATO, FÁBIO, et al. "Social CRM: Biggest Challenges to Make it Work in the Real World." International Conference on Business Information Systems. Springer, Cham, 2016.

MARCUSCHI, L. A. Análise da conversação. São Paulo: Ática, 1986.

MARTINAZZO, B., DOSCIATTI, M. M., & PARAISO, E. C. (2011). Identifying emotions in short texts for brazilian portuguese. In IV International Workshop on Web and Text Intelligence (WTI 2012) (p. 16).

MCCALLUM, ANDREW, DAYNE FREITAG, AND FERNANDO CN PEREIRA. Maximum Entropy Markov Models for Information Extraction and Segmentation. ICML. Vol. 17. 2000.

MCDONALD, R., HANNAN, K., NEYLON, T., WELLS, M. AND REYNAR, J. Structured models for fine-to-coarse sentiment analysis. In Annual Meeting-Association For Computational Linguistics (Vol. 45, No. 1, p. 432), 2007.

MEYER, D., & WIEN, F. T. (2001). Support vector machines. R News, 1(3), 23-26.

MIKOLOV, TOMAS, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

MIKOLOV, TOMAS, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

MOH, MELODY, et al. "On multi-tier sentiment analysis using supervised machine learning." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2015.

MONTGOMERY, D.C., PECK, E.A. AND VINING, G.G., Introduction to linear regression analysis. John Wiley & Sons, 2015.

MORO, S.; CORTEZ, P.; RITA, P. Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Systems with Applications, v. 42, n. 3, p. 1314-1324, 2015.

MOSQUERA, A., GUTIÉRREZ, Y., & MOREDA, P. (2017). On Evaluating the Contribution of Text Normalisation Techniques to Sentiment Analysis on Informal Web 2.0 Texts. Procesamiento del Lenguaje Natural, 58, 29-36.

MUSTO, C., SEMERARO, G., LOPS, P., & DE GEMMIS, M. (2015). CrowdPulse: A framework for real-time semantic analysis of social streams. Information Systems, 54, 127-146.

NARR, S., HULFENHAUS, M., & ALBAYRAK, S. (2012). Language-independent twitter sentiment analysis. Knowledge discovery and machine learning (KDML), LWA, 12-14.

NASCIMENTO, PAULA, et al. "Análise de sentimento de tweets com foco em notícias." Brazilian Workshop on Social Network Analysis and Mining. 2012.

NEUMANN, G., BACKOFEN, R., BAUR, J., BECKER, M. AND BRAUN, C., An information extraction core system for real world german text processing. In Proceedings of the fifth conference on Applied natural language processing (pp. 209-216). Association for Computational Linguistics, 1997.

NIELSEN, F.Å. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903, 2011.

NOPP, C., & HANBURY, A. (2015). Detecting risks in the banking system by sentiment analysis. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 591-600).

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR. Pesquisa sobre o uso das Tecnologias de Informação e Comunicação nos domicílios brasileiros – TIC Domícilios 2016. São Paulo: (NIC.br), CGI.br, 2017. ISBN 978-85-5559-048-1

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR. Pesquisa sobre o uso das Tecnologias de Informação e Comunicação nas empresas brasileiras: TIC empresas 2017. São Paulo: Comitê Gestor da Internet no Brasil, 2018. ISBN 978-85-5559-060-3. Disponível em <https://cetic.br/media/docs/publicacoes/2/TIC_Empresas_2017_livro_eletronico.pdf>

OLIVEIRA, M. A. Reanalisando o processo de cancelamento do (R) em final de sílaba. Revista de Estudos da linguagem, Belo Horizonte, v. 6, n. 2, p. 70-97, 1997.

PANG, B. AND LEE, L., Opinion mining and sentiment analysis.Foundations and trends in information retrieval, 2(1-2), pp.1-135, 2008.

PANG, B., LEE, L. AND VAITHYANATHAN, S., Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics, 2002.

PAVALANATHAN, U. AND EISENSTEIN, J. Audience-modulated variation in online social media. American Speech, 90(2), pp.187-213, 2015.

PEETERS, B. (Ed.). (2006). Semantic primes and universal grammar: Empirical evidence from the Romance languages (Vol. 81). John Benjamins Publishing.

PENNEBAKER, J.W., FRANCIS, M.E. AND BOOTH, R.J. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71, p.2001, 2001.

PETERSEN, KAI, et al. "Systematic Mapping Studies in Software Engineering." EASE. Vol. 8. 2008.

PETZ, G., KARPOWICZ, M., FÜRSCHUSS, H., AUINGER, A., STŘÍTESKÝ, V., & HOLZINGER, A. (2013). Opinion mining on the web 2.0–characteristics of user generated content and their impacts. In Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data (pp. 35-46). Springer, Berlin, Heidelberg.

PETZ, G., KARPOWICZ, M., FÜRSCHUSS, H., AUINGER, A., STŘÍTESKÝ, V., & HOLZINGER, A. (2014). Computational approaches for mining user's opinions on the Web 2.0. Infor-mation Processing & Management, 50(6), 899-908.

PIRYANI, R., MADHAVI, D., & SINGH, V. K. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. Information Processing & Management, 53(1), 122-150.

PLETSCHER-FRANKILD, S., PALLEJÀ, A., TSAFOU, K., BINDER, J. X., & JENSEN, L. J. Diseases: Text mining and data integration of disease–gene associations. Methods, 74, 83-89. 2015.

PORIA, SOUJANYA, et al. "Sentic patterns: Dependency-based rules for concept-level sentiment analysis." Knowledge-Based Systems 69 (2014): 45-63.

PRATA, DAVID N., et al. "Social Data Analysis of Brazilian's Mood from Twitter." International Journal of Social Science and Humanity 6.3 (2016): 179.].

PRICEWATERHOUSECOOPERS. They say they want a revolution – Total Retail 2016. 2016. Disponível em <https://www.pwc.com/gx/en/retail-consumer/publications/assets/total-retail-global-report.pdf >

PROFFITT ,B. "What APIs Are And Why They're Important". Disponível em: <http://readwrite.com/2013/09/19/api-defined/>. Acesso em: 12 fev. 2016

PROLLOCHS, N., FEUERRIEGEL, S., & NEUMANN, D. (2015, January). Enhancing sentiment analysis of financial news by detecting negation scopes. In 2015 48th Hawaii International Conference on System Sciences (HICSS) (pp. 959-968). IEEE.

RECUERO, R. "Redes Sociais na Internet, Difusão de Informação e Jornalismo: Elementos para discussão [Online]". Disponível em: <http://www.raquelrecuero.com/artigos/artigoredesjornalismorecuero.pdf>. Acessado em: 07 mai. 2016.

RECUERO, R. A Conversação em Rede: comunicação mediada pelo computador e redes sociais na Internet. 2 ed. Porto Alegre: Sulina, 2014b, 238 p.

RECUERO, R. Redes Sociais Na Internet. 2 ed. Porto Alegre: Sulina, 2014a, 206 p.

REZENDE, S. O. Sistemas inteligentes: fundamentos e aplicações. Barueri, Manole, 2005.

RIBEIRO, F.N., ARAÚJO, M., GONÇALVES, P., BENEVENUTO, F. AND GONÇALVES, M.A., A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods. arXiv preprint arXiv:1512.01818, 2015.

RISH, IRINA. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. IBM New York, 2001.

RIVA, H. C. (2009). Dicionário onomasiológico de expressões idiomáticas usuais na língua portuguesa no Brasil.

RODRIGUES, RAMON GOUVEIA, et al. "SentiHealth-Cancer: a sentiment analysis tool to help detecting mood of patients in online social networks." International journal of medical informatics 85.1 (2016): 80-95

SAIAS, JOSE, *et al.* "Combining overall and target oriented sentiment analysis over portuguese text from social media." Transactions on Machine Learning and Artificial Intelligence 3.3 (2015): 46.

SALTON, G., & MCGILL, M. J. (1986). Introduction to modern information retrieval.

SANTANA, V. F.; MELO-SOLARTE, D. S.; NERIS, V. P. A.; MIRANDA, L. C.; BARANUSKAS, M. C. C. Redes Sociais Online: Desafios e Possibilidades Para o Contexto Brasileiro. In: CSBC 2011. Bento Gonçalves (RS), P. 339-353, 2009.

SANTOS, CAROLINE Q., et al. "Can Visualization Techniques Help Journalists to Deepen Analysis of Twitter Data? Exploring the" Germany 7 x 1 Brazil" Case." System Sciences (HICSS), 2016 49th Hawaii International Conference on. IEEE, 2016.

SAUNDERS, C., GAMMERMAN, A., & VOVK, V. (1998). Ridge regression learning algorithm in dual variables.

SEO, J; PARK, E. A study on financing security for smartphones using text mining. Wireless Personal Communications, v. 98, n. 4, p. 3109-3127, 2018.

SEVERYN, A., & MOSCHITTI, A. (2015, august). twitter sentiment analysis with deep convolutional neural networks. in proceedings of the 38th international acm sigir conference on research and development in information retrieval (pp. 959-962). acm.

SILVA, M.J., CARVALHO, P. AND SARMENTO, L., Building a sentiment lexicon for social judgement mining. In Computational Processing of the Portuguese Language (pp. 218-228). Springer Berlin Heidelberg, 2012.

SINGH,T., & KUMARI, M. (2016). Role of Text Pre-processing in Twitter Sentiment Analysis. Procedia Computer Science, 89, 549-554.

SLOANE, D. AND MORGAN, S.P., An introduction to categorical data analysis. Annual review of sociology, pp.351-375, 1996.

SMITH, C. “By The Numbers: 170+ Amazing Twitter Statistics”. Disponível em: <http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/>. Acessado em: 14 mar. 2016

SOLAKIDIS, G. S., VAVLIAKIS, K. N., & MITKAS, P. A. (2014, August). Multilingual sentiment analysis using emoticons and keywords. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02 (pp. 102-109). IEEE Computer Society.

SOUZA, BRUNO A., *et al*. "For or against?: Polarity analysis in tweets about impeachment process of brazil president." Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web. ACM, 2016.

SOUZA, ELLEN, *et al*. "Characterizing Opinion Mining: A Systematic Mapping Study of the Portuguese Language." International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2016.

SPRINKLR. Como o social está transformando as empresas brasileiras? 2016. Disponível em <https://blog.sprinklr.com/wp-content/uploads/2016/04/20160406_WP_PT_ideas_insights_2016_parte1_web_V01.pdf>

STIILPEN JUNIOR, M., & MERSCHMANN, L. H. C. (2016, November). A Methodology to Handle Social Media Posts in Brazilian Portuguese for Text Mining Applications. In Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web (pp. 239-246). ACM.

SUN, A.; LACHANSKI, M.; FABOZZI, F. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. International Review of Financial Analysis, v. 48, p. 272-281, 2016.

SUN, S., LUO, C., & CHEN, J. (2017). A review of natural language processing techniques for opinion mining systems. Information Fusion, 36, 10-25.

TABOADA, M., BROOKE, J., TOFILOSKI, M., VOLL, K. AND STEDE, M. Lexicon-based methods for sentiment analysis. Computational linguistics,37(2), pp.267-307, 2011.+.

TELES, V., SANTOS, D., & SOUZA, E. (2016). Uma análise comparativa de técnicas supervisionadas para mineraçao de opiniao de consumidores brasileiros no twitter. Proceedings of the XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2016), 217-228.

TELLEZ, ERIC S., et al. "A case study of Spanish text transformations for twitter sentiment analysis." Expert Systems with Applications 81 (2017): 457-471. - a

TELLEZ, ERIC S., et al. "A simple approach to multilingual polarity classification in twitter." Pattern Recognition Letters 94 (2017): 68-74. - b

THE GUARDIAN. “A brief history of Facebook”. Disponível em: <https://www.theguardian.com/technology/2007/jul/25/media.newmedia>. Acessado em: 08 mai. 2016

THELWALL, M. Heart and soul: Sentiment strength detection in the social web with sentistrength. Proceedings of the CyberEmotions, pp.1-14, 2013.

THELWALL, MIKE, et al. "Sentiment strength detection in short informal text." Journal of the American Society for Information Science and Technology 61.12 (2010): 2544-2558.

TURNEY, P., Mining the web for synonyms: PMI-IR versus LSA on TOEFL, 2001.

TURNEY, P.D., Answering subcognitive turing test questions: a reply to french. Journal of Experimental & Theoretical Artificial Intelligence, 13(4), pp.409-419, 2001.

TURNEY, P.D., Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics, 2002.

TWITTER. “Twitter milestones”. Disponível em: <https://about.twitter.com/company/press/milestones>. Acessado em: 08 mai. 2016.

TWITTER. “Using hashtags on Twitter”. Disponível em: <https://support.twitter.com/articles/49309>. Acessado em: 17 fev. 2016

UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT. Information Economy Report 2017: Digitalization, Trade and Developement. Suíça, pág. 19, 2017. ISSN 2075-4396.

VAPNIK, V. Statistical learning theory. 1998. Wiley, New York, 1998.

VAPNIK, V. The nature of statistical learning theory. Springer science & business media, 1995.

VAPNIK, Vladimir N.; CHERVONENKIS, A. Ya. On the uniform convergence of relative frequencies of events to their probabilities. In: Measures of complexity. Springer, Cham, 2015. p. 11-30.

VENKATESH, V., CROTEAU, A. M., & RABAH, J. (2014, January). Perceptions of effectiveness of instructional uses of technology in higher education in an era of Web 2.0. In System Sciences (HICSS), 2014 47th Hawaii International Conference on (pp. 110-119). IEEE.

VILARES, D., THELWALL, M., & ALONSO, M. A. (2015). The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. Journal of Information Science, 41(6), 799-813.)

VITORIO; DOUGLAS, et al. "Investigating Opinion Mining through Language Varieties: a Case Study of Brazilian and European Portuguese tweets." Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology. 2017.

WASSERMAN, S. AND FAUST, K. Social network analysis: Methods and applications (Vol. 8). Cambridge university press, 1994.

WE ARE SOCIAL & HOOTSUITE. Digital in 2018: Esssential insights into Internet, Social Media, Mobile, and Ecommerce use around the world. Disponível em < https://wearesocial.com/blog/2018/01/global-digital-report-2018>

XU, KAIQUAN, et al. "Mining comparative opinions from customer reviews for Competitive Intelligence." Decision support systems 50.4 (2011): 743-754.

YEE LIAU, B.; PEI TAN, P. Gaining customer knowledge in low cost airlines through text mining. Industrial Management & Data Systems, v. 114, n. 9, p. 1344-1359, 2014.

YU, Y., DUAN, W., & CAO, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. Decision Support Systems, 55(4), 919-926.

ZEPHORIA. “The Top 20 Valuable Facebook Statistics – Updated July 2018”. Disponível em <https://zephoria.com/top-15-valuable-facebook-statistics/>. Acessado em: 30 julho. 2018.

ZHANG, X., & LECUN, Y. (2015). Text understanding from scratch. arXiv preprint arXiv:1502.01710.