DICIONÁRIO DE POLARIDADES PARA APOIO A ANÁLISE DE SENTIMENTO

Paula Camargo Nascimento

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Setembro de 2014

DICIONÁRIO DE POLARIDADES PARA APOIO A ANÁLISE DE SENTIMENTO

Paula Camargo Nascimento

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Prof. Geraldo Bonorino Xexéo, D.Sc.

_____

Prof. Jano Moreira de Souza, Ph.D.

_____

Prof. Adriana Santarosa Vivacqua, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

SETEMBRO DE 2014

Nascimento, Paula Camargo

Dicionário de Polaridades para Apoio a Análise de Sentimento / Paula Camargo Nascimento. – Rio de Janeiro: UFRJ/COPPE, 2014.

XV, 93 p.: il.; 29,7 cm.

Orientador: Geraldo Bonorino Xexéo.

Dissertação (Mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2014.

Referências Bibliográficas: p. 75-81.

1. Criação de Dicionários de Polaridades. 2. Análise de Sentimento. 3. Similaridade Sintagmática de Termos. 4. Classificação da Orientação de Termos. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

À minha avó Dalva (in memoriam).

Agradecimentos

Agradeço aos meus pais, Lilian e Valter, por tudo que eu sou, tudo que eu tenho e pela oportunidade de sempre ter seguido por onde meus pés quisessem me levar com o conforto de saber que sempre me apoiariam. Muito obrigada por me permitirem chegar até aqui. Não tenho dúvidas de que os principais responsáveis por esta conquista são vocês.

Ao meu noivo, amigo e companheiro Rodrigo, por acreditar em mim, até quando eu mesma não acredito, e por ser um dos maiores incentivadores que eu tenho em tudo que eu faço, sempre.

Ao meu orientador Geraldo Bonorino Xexéo por todo o apoio e confiança ao longo desta jornada e por todo o suporte fundamental para a conclusão deste trabalho.

Agradeço também aos professores Jano Moreira de Souza e Adriana Santarosa Vivacqua por aceitarem participar da banca de defesa de mestrado.

Aos amigos da GPE – Gestão de Processos Estratégicos por todo o apoio e compreensão durante o período de desenvolvimento deste trabalho. Em especial, agradeço ao Bruno Osiek por todo o conhecimento repassado durante diversas discussões sobre o assunto, pelos conselhos e principalmente por ter me apresentado a área de Recuperação da Informação ainda durante meu curso de graduação. Também faço um agradecimento especial à Marcelo Areas, Fabrício Pereira, Caio Ribeiro, André Moreira, Leonardo Marques, Ricardo Barros e Ester Lima, que me ofereceram todo o suporte e tranquilidade para que eu pudesse me dedicar à este trabalho.

À todos os meus amigos e familiares por aguentaram firme e com paciência os momentos de ausência e de estudo. Em especial, agradeço ao Fabio Galluzzo, Júlia Almeida e Rafael Espirito Santo por toda a ajuda durante este período.

Por fim, agradeço a Deus e ao fato de Ele colocar no meu caminho todas as oportunidades e obstáculos necessários para que eu pudesse realizar este trabalho.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

DICIONÁRIO DE POLARIDADES PARA APOIO A ANÁLISE DE SENTIMENTO

Paula Camargo Nascimento

Setembro/2014

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Esta dissertação apresenta uma estratégia de criação de um dicionário de polaridades para apoiar as iniciativas de Análise de Sentimento de textos escritos em Português do Brasil. Para isso, serão analisados diferentes métodos de classificação da orientação de termos de forma a eleger o que melhor se adequa a esta tarefa, além de permitir comparar o desempenho entre esses e outros métodos de combinação de resultados aqui explorados. Além disso, a estratégia aqui proposta também considera a inclusão de termos extraídos a partir dos documentos analisados, adicionando conhecimento específico do domínio tratado ao dicionário final gerado. Como todo o processo para a criação do dicionário final será apresentado, este trabalho também tem a intenção de repassar o conhecimento adquirido para que a geração de outros dicionários, com base em assuntos diferentes do analisado neste trabalho, seja possível.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

DICTIONARY OF POLARITIES FOR SUPPORTING SENTIMENT ANALYSIS

Paula Camargo Nascimento

September/2014

Advisor: Geraldo Bonorino Xexéo

Department: Computer Science and Systems Engineering

This paper presents a strategy for creating a dictionary of polarities to support initiatives of Sentiment Analysis in Portuguese from Brazil. For this, different methods of polarity classification of terms will be analyzed in order to elect which one suits best to this task. It will also allow a performance comparison between these and other methods of results ensemble, also explored in this work. Moreover, the strategy proposed here considers the inclusion of terms extracted from the corpus, adding specific domain knowledge to the lexicon generated at the end. As the whole process for creating the final dictionary will be presented, this work also intends to transmit the acquired knowledge, making the generation of other dictionaries, based on different domains than those analyzed in this work possible.

Índice

Índice de Figuras

Índice de Tabelas

1 Introdução

1.1 Motivação

Saber o que as pessoas pensam ou como elas se sentem sempre foi objeto de interesse (Pang & Lee, 2008). Seja para decidir qual roupa iremos a um evento ou qual carro compraremos, considerar a opinião daqueles que já passaram por experiência similar sempre foi um dos fatores mais importantes para chegarmos à conclusão de qual caminho seguir. Com a popularização da Internet e o surgimento da Web 2.0, nossas fontes de opinião deixaram de estar limitadas a conhecidos, parentes e amigos e passaram a incluir também a multidão que navega na Internet e o enorme volume de conteúdo subjetivo – textos que expressam sentimento e opinião sobre alguma entidade (Liu, 2010) – gerado diariamente (Pang & Lee, 2008). Esta nova forma de expor opiniões e sentimentos trouxe à tona um novo comportamento da população: as pessoas passaram a ter maior confiança em produtos e serviços que tivessem sido bem avaliados em plataformas *online* e passaram a considerar a inteligência coletiva – termo que define a situação em que grupos de pessoas, cada uma com sua experiência individual, geram conhecimento (Malone et al, 2009) – como um dos insumos mais importantes do seu processo de tomada de decisão (Pang & Lee, 2008).

Essa nova atitude por parte dos consumidores fez com que a indústria demonstrasse interesse na análise das opiniões divulgadas na Internet devido ao alto potencial de influência dessas informações. Aliado a isto, a partir do ano de 2001 aproximadamente, novas tecnologias e possibilidades apareceram, tais como "os métodos de aprendizagem de máquina para processamento de linguagem natural" (Pang & Lee, 2008) e a disponibilidade de grandes volumes de dados para serem analisados. A lista de desafios oferecidos para os pesquisadores dispostos a se envolver na tarefa de analisar esse conteúdo, e extrair as opiniões lá descritas, completa o conjunto dos mais significantes fatores que desencadearam a ascensão da área de estudo hoje conhecida por Análise de Sentimento (Pang & Lee, 2008). Tal ascensão define um marco no campo de Extração da Informação, onde o foco deixa de ser apenas o lado objetivo do texto, ou seja, o fato, para passar a tratar também o lado subjetivo, ou seja, a opinião (Liu, 2010).

Neste contexto, a partir de 2006, popularizou-se na Internet uma nova forma de comunicação, conhecida como *microblogging*. Tal conceito ganhou destaque através da

rede social Twitter (Honeycutt & Herring, 2009), que tinha como proposta inicial permitir que seus usuários informassem à sua rede de amigos o que eles estavam fazendo, utilizando apenas 140 caracteres e em tempo real (140 Characters, 2009). Com o passar dos anos, a união dessas características com o elevado nível de popularidade alcançado por esta ferramenta e o uso dado a ela por seus usuários, fez com que o Twitter passasse a ser considerado não mais uma rede social, mas sim uma rede de informações (Twitter, 2012) recheada com as opiniões e sentimentos das pessoas que dela participam (Naaman & Boase, 2010). No Brasil, isso representa ter acesso à opinião de cerca de 41.200.000 pessoas que, em junho de 2012, somando apenas São Paulo e Rio de Janeiro, geraram aproximadamente 2,3 bilhões de mensagens, conhecidas como *tweets*, exprimindo seus sentimentos sobre diversos acontecimentos e assuntos (Semiocast, 2012).

Em 2010, um estudo realizado por (Kwak et al, 2010) visava avaliar justamente o potencial do Twitter como uma rede de informação. Tal estudo mostrou que cerca de 85% dos *tweets* gerados estão relacionados a manchetes de jornal e demais notícias divulgadas na mídia. Além disso, em 2011, um trabalho cujo foco era classificar o sentimento – identificar se o texto exprime raiva, felicidade, alegria ou tristeza, por exemplo – presente em *tweets* foi realizado por (Bollen et al, 2011) e nele foi constatado o enorme potencial do Twitter para auxiliar na tarefa de analisar o sentimento da população. Estes resultados tornam o Twitter o candidato ideal para validar este trabalho cujo objetivo é criar um dicionário de polaridades em português, que permita melhorar os resultados obtidos com a classificação automática de sentimentos presentes em textos deste idioma.

Aliado a isto, em (Li & Li, 2011) foi constatado ainda que 20% dos *tweets* estão relacionados a marcas e expressam opinião sobre "a empresa, produto ou serviço". Além disso, uma pesquisa realizada por (Stelzner, 2012) mostra que 83% dos profissionais de marketing veem as mídias sociais como ferramentas importantes para as empresas em que trabalham e o Twitter está em segundo lugar entre essas mídias, ficando atrás apenas do Facebook, outra rede social extremamente utilizada atualmente por usuários de todo o mundo. Sendo assim, analisar informações divulgadas no Twitter se tornou, também, uma necessidade estratégica das empresas, devido ao alto número de pessoas presentes nessas redes, sendo influenciadas diariamente pela opinião de seus contatos (Li & Li, 2011). Para esses profissionais e para toda a população, seguindo nossos instintos naturais, a análise da orientação – ou seja, identificar se as opiniões são positivas, negativas ou neutras – das mensagens publicadas em serviços de *microblogging* vão ao encontro às

necessidades de saber o que as pessoas pensam e como elas reagem ao que acontece no dia a dia.

1.2 Problema

Na seção anterior foi citado que um dos principais motivos para a ascensão da área de estudo foco deste trabalho foram os desafios oferecidos por ela. Dentre eles estão: verificar se o texto analisado é relevante para o tópico de interesse; verificar se o texto é subjetivo; verificar qual é o sentimento expresso neste texto; e, por fim, identificar a melhor forma de sumarizar as informações coletadas (Pang & Lee, 2008). Este trabalho foca nestas questões, principalmente na análise do sentimento contido no documento analisado.

Este desafio já foi amplamente explorado em diversos domínios e através de diversas estratégias, principalmente se considerarmos outros idiomas, como a língua inglesa. No entanto, se considerarmos a análise de textos escritos em português, identificamos poucos trabalhos na área, destacando-se entre eles (Silva et al, 2010), cujo objetivo foi construir um dicionário de polaridades com termos capazes de classificar pessoas, e (Souza et al, 2011), que gerou um dicionário genérico a partir da união de três estratégias: tradução de termos, incorporação do thesaurus TeP 2.0 (Maziero et al, 2008) e aplicação do algoritmo de similaridade de termos por coocorrência apresentado por (Turney, 2002). Além destes, podemos destacar também o trabalho desenvolvido em (Carvalho et al, 2009), com o objetivo de identificar frases que indicavam ironias em textos extraídos de comentários de um jornal brasileiro *online*.

Esta escassez de trabalhos com foco na língua portuguesa talvez seja uma consequência do fato de que importantes recursos sejam construídos e melhorados apenas para outros idiomas, dificultando assim a análise de textos em português tomando como base resultados de trabalhos já realizados previamente. Além disso, quando os documentos tratados são gerados a partir de serviços de *microblogging*, novos desafios são apresentados devido às suas características, já anteriormente citadas, que geram textos escritos de modo extremamente informal e inconsistente (Brew et al, 2011). Em muitos casos retirados do Twitter, a sentença apresenta diversas gírias, siglas e expressões próprias dos usuários, dificultando ainda mais o processo de análise do sentimento exposto no texto, especialmente quando termos específicos do domínio não são considerados.

Devido à alta popularidade do Twitter e ao seu potencial como gerador de documentos subjetivos (Naaman & Boase, 2010), diversos estudos e técnicas de análise de *tweets* na língua inglesa já foram desenvolvidos e testados. No entanto, não conseguimos identificar um número significativo de iniciativas de análise de sentimento desses textos em português. Dada a alta participação de usuários brasileiros nessa rede, acredita-se que analisar o conteúdo gerado por eles trará informações bastante valiosas sobre o que as pessoas pensam em relação ao que acontece e ao que é produzido dentro do país.

1.3 Objetivo

O objetivo deste trabalho é construir e avaliar um dicionário de polaridades adaptado ao domínio dos documentos analisados, visando apoiar a avaliação automática de sentimentos presentes em textos escritos em português. Como o processo de construção do léxico em questão será apresentado desde o princípio, este trabalho também tem por objetivo servir como base para a criação de dicionários de polaridades na língua portuguesa, focados em outros domínios diferentes do tratado aqui. Através deste trabalho, espera-se gerar um importante artefato que seja capaz de aumentar a qualidade dos resultados obtidos com as tecnologias de classificação de sentimentos disponibilizadas atualmente para o português e também superar uma deficiência de diversos métodos baseados na criação de dicionários, que é ser independente de domínio. Tal característica é considerada uma deficiência já que domínios distintos podem oferecer polaridades distintas para uma mesma palavra, devido ao significado dado a ela em determinado tema. (Qiu et al, 2009).

1.4 Método proposto

O método proposto neste trabalho tem como base a estratégia de criação de dicionários utilizada em (Silva et al, 2010), onde a junção de dicionários pré-existentes e a extração de palavras do *corpus* permitiu a criação de um novo dicionário com foco em palavras capazes de classificar pessoas. Sendo assim, para alcançar os objetivos descritos anteriormente, o método proposto neste trabalho segue três etapas principais, são elas:

1. Criação de um dicionário inicial, sem influência do domínio analisado, a partir da união de léxicos gerados anteriormente por (Maziero et al, 2008) e por (Silva et al, 2010). Esta etapa permitiu obter um dicionário composto por 114.268 termos, com relações de sinonímia e antonímia, extraídas do dicionário gerado

anteriormente por (Maziero et al, 2008), entre cerca de 39% deles. Além disso, 74.964 termos já possuíam polaridade obtida através do trabalho desenvolvido por (Silva et al, 2010).

2. Inclusão de novos termos no dicionário, obtidos através da aplicação de regras de extração de palavras do texto. Tais regras foram propostas em (Silva et al, 2010) e generalizadas e adaptadas para as estruturas textuais consideradas neste trabalho. Esta etapa permitiu que novos termos relacionados ao domínio tratado nos documentos analisados fossem incluídos no dicionário.
3. Cálculo e validação das polaridades dos demais termos do dicionário. Este cálculo se deu através de diferentes estratégias, considerando as relações de sinonímia e antonímia previamente estabelecidas entre os termos considerados e também cálculos de similaridade entre as palavras dos textos analisados, como veremos mais adiante.

É importante ressaltar que este trabalho se diferencia de (Silva et al, 2010) por ser adaptável ao domínio dos documentos analisados e não manter seu foco em apenas um domínio. Além disso, diferentemente de (Silva et al, 2010), este trabalho ainda realiza testes com diferentes estratégias de cálculo de polaridades para definir, entre elas, qual possui maior qualidade de resultados para tal tarefa.

Ao longo deste trabalho foi possível identificar apenas uma iniciativa com metodologia bastante similar à utilizada neste estudo e aplicada para o português, realizada em (Santos et al, 2011). No entanto, este trabalho se diferencia de (Santos et al, 2011) por também tratar termos que foram extraídos dos documentos analisados e, portanto, são específicos do domínio explorado.

1.5 Organização

Além deste capítulo de introdução, este trabalho possui mais cinco capítulos. No capítulo 2, os conceitos de *microblogging* e estudos relacionados ao Twitter são explorados. No Capítulo 3, é apresentada uma revisão bibliográfica da área de análise de sentimento. No capítulo 4, é apresentado o método para a geração de dicionários de polaridades utilizado neste trabalho. No capítulo 5, são apresentados os resultados encontrados e, por fim, o capítulo 6 apresenta as conclusões, discute as limitações existentes e apresenta possíveis trabalhos futuros.

2 Twitter: uma rede de informações

2.1 Introdução

Conforme dito anteriormente, a partir de 2006, os serviços de *microblogging* se popularizaram entre aqueles que faziam uso da Internet na época. Antes deles, a maneira mais conhecida de se publicar na rede o que se queria dizer era através dos *blogs* (Honeycutt & Herring, 2009). Apesar de existirem no formato atual desde 1996, os *blogs* se tornaram mais conhecidos a partir de 1999 (Herring et al, 2004) e até hoje o número de *blogs* na Internet continua a crescer, como mostra um estudo realizado pela NM Incite que verificou que, entre 2006 e 2011, mais de 137.228.000 novos *blogs* foram criados (Nielsenwire, 2012). No entanto, os novos serviços de *microblogging* permitiram que os blogueiros – pessoas que divulgam suas opiniões através de *blogs* – diminuíssem o tempo e o esforço gastos na criação de uma publicação tradicional, passando a gerar diversas pequenas publicações diariamente (Java et al, 2007).

Um *microblog* nada mais é do que um novo tipo de *blog* onde as pessoas se expressam em um número limitado de palavras, geralmente até 140 caracteres, e compartilham suas opiniões com o público em geral ou com um grupo restrito de contatos. Inicialmente, aqueles que utilizavam serviços de *microblogging* tinham por objetivo divulgar aquilo que estavam fazendo, razão pela qual o primeiro slogan do Twitter, ferramenta mais popular de *microblogging*, continha a frase "O que você está fazendo?" (Honeycutt & Herring, 2009). Com o passar dos anos e com o aumento de popularidade entre os usuários, os assuntos tratados nestas publicações passaram a conter desde opiniões sobre diversos acontecimentos no mundo até conteúdo promocional de serviços e produtos, além do seu enfoque inicial.

A adoção em massa dos *microblogs* trouxe uma mudança na maneira como as pessoas produzem e consomem conteúdo na Internet. Segundo (Ebner & Schiefner, 2008), o sucesso dos *blogs* se deu, principalmente, devido ao fato de serem fáceis de usar, conectarem pessoas com interesses similares e compartilharem sentimentos e pensamentos próprios de uma pessoa, o dono do *blog*. Já o sucesso do *microblog*, acredita-se, deu-se por proporcionar às pessoas uma forma de gerar conteúdo com rapidez e dinamismo. Devido à publicação dos textos imediatamente após serem escritos, de qualquer lugar, e por serem curtos e objetivos, as pessoas passaram a expor mais seus

sentimentos e pensamentos em relação a tudo que acontece a sua volta, no momento em que o evento ocorre. Essa nova atitude gerou nas pessoas o comportamento de se manterem atualizadas, sobre a vida pessoal ou não dos demais, através dos *microblogs*, lendo diversas vezes por dia, também de forma rápida e dinâmica, as milhares de publicações geradas por seus contatos.

Um fator muito importante para toda essa mudança de paradigma foi a popularidade alcançada pelo Twitter. Criado com o objetivo de oferecer uma maneira simples das pessoas se comunicarem sobre suas atividades diárias, em pouco tempo passou a agregar opiniões de pessoas por todo o mundo. O número de usuários que esta ferramenta possui e o uso que essas pessoas dão a ela fizeram com que o Twitter deixasse de ser uma rede social para passar a ser uma rede de informações, gerando conteúdo rico em subjetividade sobre qualquer evento que acontece nos quatro cantos do mundo (Naaman & Boase, 2010). Por esta razão, este capítulo se dedica a apresentar o Twitter, tendo como foco a sua aplicação acadêmica. Para isto, a seção 2.2 explica o funcionamento do Twitter e sua estrutura. A seção 2.3 trata dos estudos relacionados a entender melhor as razões que motivam o uso desta ferramenta e como é o comportamento daqueles que a utilizam. A seção 2.4 traz a visão do conteúdo que é gerado no Twitter e quais são as aplicações atuais desses dados. E, por fim, a seção 2.5 conclui este capítulo.

2.2 O que é o Twitter?

A atual auto definição do Twitter diz que ele é uma rede de informação e "o meio mais rápido e fácil de se manter perto de tudo que te interessa" (Twitter, 2012). Esta afirmação justifica-se principalmente pelo formato dessa ferramenta, que baseia-se na publicação de mensagens, de até 140 caracteres, denominadas *tweets*. Os *tweets* permitem que as pessoas compartilhem suas "... últimas histórias, ideias, opiniões e notícias" em tempo real, de maneira pública ou apenas para um grupo selecionado de pessoas (Twitter, 2012). Esta característica define o conceito de *microblog*.

Por ter surgido como uma rede social, o Twitter mantém algumas características comuns a esses sites, como a criação de um perfil e a formação de redes com outros usuários. No entanto, a formação de redes nesta plataforma possui um comportamento um pouco diferente das demais plataformas sociais. No Twitter, existem dois conceitos de relacionamento: seguir e ser seguido. As pessoas seguem um usuário com o intuito de

ler suas informações, e são seguidas porque alguém teve esse mesmo interesse em relação a elas. Esse relacionamento não precisa ser recíproco, ou seja, quem é seguido por uma pessoa não precisa segui-la de volta e esta é uma das principais características que define a dinâmica de uso do Twitter, como veremos adiante (Twitter, 2013a).

Outro diferencial do Twitter em relação a outras redes sociais é o foco no conteúdo gerado. As pessoas estão no Twitter para observar publicações com informações interessantes à elas, sejam essas sobre assuntos públicos ou pessoais. Por isso, diversas funcionalidades foram desenvolvidas para melhorar a comunicação entre as pessoas. Citaremos aqui quatro desses mecanismos que consideramos mais relevantes para este trabalho, são eles (Twitter, 2013b):

1. *Retweets*: Representam a ação de um usuário repassar aos seus seguidores um *tweet* lido por ele, por tê-lo considerado interessante o suficiente para ser compartilhado. Representam também o *tweet* repassado em si.
2. Menções: São representadas por '@usuario' e significam mencionar alguém em uma mensagem. Essas menções são apresentadas nos *tweets* como *links* para os perfis correspondentes, onde estarão listadas todas as publicações daquele usuário, caso seu perfil seja público.
3. *Hashtags*: Criada pelos próprios usuários, as *hashtags* indicam o tópico ou as palavras-chave do *tweet* que as contém. São identificadas pelo símbolo '#' e comumente se tornam *trends*, que serão explicadas a seguir. Elas também são apresentadas como *links*, que seguem para uma listagem de *tweets* que carregam a mesma *hashtag*, ou seja, falam do mesmo assunto.
4. *Trends*: Representam os tópicos mais populares no Twitter, separados por região. Estes tópicos são apresentados para os usuários em forma de listagem, que permite o acesso aos *tweets* relacionados.

Por último, o Twitter possui um vocabulário próprio. Isso se deve principalmente ao fato de as pessoas precisarem se expressar com uma grande limitação de caracteres. Por esse motivo, diversas abreviações e gírias próprias da plataforma foram criadas pelos usuários ao longo do tempo, o que deu ao Twitter a característica de ter um linguajar diferenciado em seu conteúdo.

2.3 Por que as pessoas “tweetam”?

O título desta seção se inspira no título de um dos primeiros trabalhos acadêmicos cujo foco era o Twitter, realizado em (Java et al, 2007). No início, o objetivo principal da academia era entender como e por que as pessoas usavam o Twitter e, no trabalho realizado por (Java et al, 2007), foi constatado que, na época, as pessoas utilizavam o Twitter para comentar suas rotinas diárias, conversar com amigos, compartilhar informações e reportar notícias, nesta ordem. Além disso, os usuários se encaixavam em três grupos principais: as fontes de informação, os amigos e aqueles que buscavam informação.

De 2007 para cá, pouca coisa mudou no que diz respeito ao modo como podemos classificar os grupos de usuários existentes no Twitter, conforme mostrado em (Xu et al, 2012). No entanto, o uso dado à ferramenta foi modificado ao longo dos anos e fez com que a rede passasse a tomar um novo rumo em relação ao conteúdo gerado por seus usuários. Conforme citado anteriormente, em (Kwak et al, 2010), foi visto que o conteúdo gerado em relação às notícias divulgadas na mídia passou a ser maioria entre os tópicos falados na rede, deixando para trás o objetivo inicial do Twitter. A seguir, tentaremos entender, através dos trabalhos realizados na área, como e por que essa evolução ocorreu.

2.3.1 A motivação por trás do uso do *microblog*

Para entendermos esse processo de evolução dos assuntos comentados no Twitter é preciso, primeiramente, buscar os motivos que levam as pessoas a divulgarem informações em plataformas de *microblog* e a quem elas estão conectadas.

Em (Zhao & Rosson, 2009), uma análise foi conduzida para avaliar, três anos após o surgimento do Twitter, qual o papel do *microblog* na comunicação informal entre as pessoas. Neste estudo, foi possível identificar a motivação por trás dos três grupos de usuários identificados no trabalho de (Java et al, 2007), e confirmados em (Krishnamurthy et al, 2008). O primeiro deles é o grupo de pessoas que busca, através das redes sociais, estabelecer laços com aqueles que, por algum motivo, estão distantes fisicamente e não fazem parte das atividades do seu dia a dia. Essas ligações podem ser estabelecidas com amigos, colegas de trabalho ou pessoas que não se conhecem pessoalmente, mas que compartilham os mesmos interesses. O fato importante é que, nesses casos, as plataformas de *microblog* desempenham o papel de substitutos das conversas informais que ocorrem

quando encontramos alguém que há muito não vemos. Seguir alguém no Twitter, significa ter acesso aos seus interesses e aos acontecimentos mais importantes de sua vida, o que cria uma sensação de proximidade, mesmo que virtual. Tais constatações vão ao encontro ao estudo realizado em (Huberman et al, 2008) que diz que apesar da enorme quantidade de usuários conectados à rede de uma pessoa, ela se comunica apenas com uma pequena parte dela, apenas com aqueles que realmente interessam.

Segundo (Zhao & Rosson, 2009), três características próprias do *microblogging* contribuem para que o Twitter seja visto desta forma. A primeira está ligada à estrutura de uso do *microblog* que permite que as pessoas divulguem de forma rápida e fácil qualquer informação que julguem ser interessante o suficiente para ser divulgada com os demais. Sendo assim, os usuários publicam diversos acontecimentos diários apenas por acharem pouco custoso escrever frases de no máximo 140 caracteres e dividir isso com seus seguidores através do Twitter. A segunda característica é a oportunidade de divulgar uma informação em tempo real. Esta característica estimula a conversação e torna o Twitter uma fonte dos últimos acontecimentos, sejam eles pessoais ou não. A união destes dois fatores dá origem à terceira característica citada por (Zhao & Rosson, 2009): o Twitter é um *feed* de notícias baseado em pessoas. Como exemplo dessa constatação, temos o perfil @LeiSecaRJ. Atualmente, este perfil é utilizado como fonte de informações sobre o trânsito na cidade do Rio de Janeiro em tempo real e todos os dados publicados são gerados pelos próprios usuários do Twitter e seguidores do @LeiSecaRJ.

Ao aliar a característica de *feed* à enorme quantidade de usuários e mensagens diárias publicadas, o Twitter virou uma fonte muito rica para aqueles que buscam informação, e uma oportunidade de atingir milhões de pessoas para aqueles que buscam divulgar informações – os outros dois perfis de usuários encontrados em (Java et al, 2007). A união desses interesses fez com que, ao passar dos anos, os usuários passassem a ver o Twitter como um meio de se manter atualizado sem precisar buscar pelas informações, apenas esperando que elas venham até eles.

2.3.2 Os relacionamentos entre os usuários

A partir de 2010, é nítida a alteração do foco de estudo sobre o Twitter. Até então, a maioria dos trabalhos buscava entender a dinâmica desta nova forma de comunicação e o que estimulava as pessoas a fazerem parte desta rede. Após este primeiro momento, a curiosidade dos pesquisadores se voltou para os usuários e os seus relacionamentos.

Na subseção anterior, classificamos os grupos de usuários identificados em (Java et al, 2007) de acordo com suas motivações para usar o Twitter. No entanto, há uma outra perspectiva de caracterização desses grupos, ligada a quantidade e ao tipo de relacionamentos que eles possuem (Java et al, 2007), (Krishnamurthy et al, 2008).

O primeiro deles, conhecido como Fontes de informação (Java et al, 2007) ou Emissores (Krishnamurthy et al, 2008), é composto por usuários que possuem enorme quantidade de usuários seguidores se comparado ao número de usuários que eles mesmos seguem. Muitos desses usuários são representantes de alguma mídia ou empresa que faz uso do Twitter para divulgar suas informações, campanhas e promoções para o público (Chu et al, 2010). Atualmente, muitas dessas empresas fazem uso de programas definidos em (Chu et al, 2010) como *cyborgs*, ou seja, programas que auxiliam ou são auxiliados por humanos para divulgarem, automaticamente e periodicamente, *tweets* com informações e propagandas. Além desses, também existem os *bots*, que são programas capazes de gerar esse mesmo conteúdo sem a necessidade de interação humana.

O segundo grupo, conhecido como Amigos (Java et al, 2007) ou Conhecidos (Krishnamurthy et al, 2008), é composto por usuários que tendem a apresentar um equilíbrio entre a quantidade de seguidores que possuem e o número de usuários que seguem. Acreditamos que este é o grupo de usuários que corresponde ao comportamento observado por (Mischaud, 2007) e citado por (Honeycutt & Herring, 2009), onde quase 60% das mensagens trocadas no Twitter estabeleciam algum tipo de conversação. Relacionando estes dados aos já apresentados, este é o grupo de pessoas que busca interagir com pessoas distantes e que, portanto, estabelece uma rede apenas com aqueles com quem realmente querem trocar informações e estabelecer um contato maior através do microblog. Segundo (Honeycutt & Herring, 2009), esta característica viabiliza a utilização do Twitter para fins de colaboração, assim como observado por (Dunlap & Lowenthal, 2009), que mostrou ser possível utilizar a plataforma para fins de colaboração na educação.

O terceiro e último grupo corresponde aos Buscadores de informação (Java et al, 2007), pessoas que seguem muito mais usuários do que o número de seguidores que possui. Para (Krishnamurthy et al, 2008), essas pessoas se caracterizam como *Spammers*, por seguirem o maior número de usuários possível com a intenção de que parte deles passem a segui-lo para que ele possa divulgar seus *tweets* para uma quantidade de pessoas

cada vez maior. No entanto, esta definição de *Spammer* não é apoiada em (Benevenuto et al, 2010) já que, no Twitter, não há nenhuma obrigação em seguir algum usuário. Segundo (Benevenuto et al, 2010), *Spammers* são aqueles que buscam se infiltrar no Twitter através da divulgação de mensagens relacionadas a assuntos altamente comentados e buscados pelos demais, fornecendo links para conteúdos maliciosos. Relacionando ao trabalho realizado em (Chu et al, 2010), a necessidade deste perfil de usuário deu origem aos chamados *bots* maliciosos, utilizados para divulgar *spams* e conteúdo malicioso constantemente na rede.

2.3.3 A evolução do conteúdo gerado

A junção da utilização do Twitter como um *feed* baseado em pessoas com a enorme quantidade de usuários da plataforma propiciou a alteração do uso principal dado ao *microblog*. Os tópicos que antes apareciam em último na lista de atividades exercidas no Twitter (Java et al, 2007), passaram a aparecer em primeiro. Isso mostra que, atualmente, o Twitter é de fato uma rede de informações sobre tudo que acontece no mundo, gerando conteúdo mais relacionado à notícias públicas do que a acontecimentos pessoais de seus usuários (Kwak et al, 2010).

Alguns fatores contribuíram para esta transformação de conteúdo. O primeiro deles, e possivelmente o mais decisivo, é a facilidade de divulgar informações no momento em que temos contato com elas. Essa facilidade deve-se, principalmente, à popularização de celulares que disponibilizam o uso da Internet e à enorme variedade de maneiras de publicar conteúdo no Twitter através desses aparelhos (Castillo et al, 2011). Esta característica permite, por exemplo, que uma pessoa divulgue que há um acidente em uma via ao passar por ali ou que as pessoas comentem ou divulguem para as outras alguma comunicação oficial das autoridades. Segundo (Castillo et al, 2011) esse diferencial do Twitter faz com que "ele seja um ambiente ideal para a disseminação de notícias de última hora direto da fonte jornalística ou do local em que a notícia ocorreu".

Ao perceber este enorme potencial de comunicação que as ferramentas de *microblog* oferecem, essas plataformas passaram a ser alvo de outros tipos de usuários, que também contribuíram para a alteração do uso da ferramenta. Esses usuários podem ser classificados como Emissores e representam, por exemplo, empresas e mídias de massa, como rádios e emissoras de TV. Essas empresas perceberam o potencial das redes sociais, e principalmente do Twitter, como nova mídia de comunicação com o público e

passaram a incluí-las em suas políticas de marketing (Chu et al, 2010). Conforme citado anteriormente, os dados levantados em (Stelzner, 2012) comprovam este comportamento e mostram que esta mídia possui lugar de extrema importância para as empresas atualmente.

O fato de esses dois elementos serem motivadores da mudança de conteúdo do Twitter é fortalecido por dois fatores comportamentais: a credibilidade e a influência entre usuários. Segundo estudo realizado em (Castillo et al, 2011), a credibilidade no Twitter está ligada à veracidade de uma publicação, ou seja, é ser possível saber se os usuários vão acreditar ou não em uma informação divulgada no *microblog*. Este estudo verificou que a credibilidade aumenta quando a notícia é publicada por um usuário já reconhecido na rede, com alto nível de influência – onde ser influente no Twitter significa demonstrar aos seus contatos ser participativo e envolvido com o assunto em questão (Cha et al, 2010). Para apoiar a importância destes fatores comportamentais está o fato de as pessoas se motivarem a divulgar informações que podem ajudar aos outros, por acreditarem estar contribuindo para a comunidade ou para ser reconhecido (Malone et al, 2009). Esta característica humana estimula a participação das pessoas e propicia o aumento de publicações informativas consideradas verídicas, fortalecendo ainda mais o papel do Twitter como rede de informações.

2.4 O que as pessoas “tweetam”?

Uma outra perspectiva dos trabalhos acadêmicos que exploram o Twitter são as aplicações dadas às informações presentes nos *tweets* publicados. Como vimos anteriormente, com o passar dos anos, o uso dado ao Twitter e, por consequência, o conteúdo publicado nessa rede foi mudando. Esta transformação aliada ao aumento da população do Twitter fez com que o conteúdo publicado pelos usuários desta rede passasse a ganhar notoriedade e ser aplicado em diversas áreas de análise de textos, com diferentes objetivos, sendo os principais listados abaixo.

2.4.1 Detecção e acompanhamento de eventos em tempo real

Uma das aplicações mais exploradas é a detecção e acompanhamento de eventos em tempo real. A motivação por trás dessa vertente de pesquisa está nas possibilidades oferecidas pelo próprio formato das ferramentas de *microblogging*. Quando algum fato

inesperado acontece, é possível perceber o aumento de mensagens relacionadas e assim detectar um novo acontecimento ou comportamento da população (Earle et al, 2011).

Um dos primeiros trabalhos a explorar essa característica do Twitter foi (Sakaki et al, 2010). Assim como em (Achrekar et al, 2011), (Earle et al, 2011), (Culotta, 2010) e (Vieweg et al, 2010), o objetivo era analisar o conteúdo divulgado no Twitter para detectar a ocorrência de situações emergenciais e conseguir prever, antes das autoridades, localidades que precisariam de assistência médica ou resgate de vítimas. Já em (Cataldi et al, 2010), (Mathioudakis & Koudas, 2010), (Petrovic et al, 2010) e (Phuvipadawat & Murata, 2010), o objetivo era utilizar os *tweets* gerados pela população para descobrir notícias de última hora, antes que as grandes mídias jornalísticas tomassem conhecimento. Por fim, em (Conover et al, 2011), (Metaxas et al, 2011) e (Younus et al, 2011), as opiniões expressas nas mensagens publicadas no Twitter foram utilizadas para avaliar a opinião pública sobre políticos e tentar prever o resultado de eleições.

Todos esses estudos fazem uso da capacidade do Twitter em gerar grande volume de conteúdo em um curto espaço de tempo e sempre muito próximo do instante e local em que o evento tratado na mensagem ocorreu. Mais do que nunca, nesses casos, o Twitter é utilizado como um *feed* baseado em pessoas e os textos publicados por elas são analisados para extrair eventos no momento em que eles ocorrem.

2.4.2 Análise de Sentimento de *tweets*

Outra aplicação amplamente explorada é a análise do sentimento existente nos *tweets* publicados. Esta aplicação tem como principais motivações o volume de conteúdo gerado e a alta carga subjetiva dos textos, conforme citado anteriormente. Neste caso, a análise de sentimento tem fins não apenas acadêmicos, mas também de mercado, ao ser utilizado para detectar a opinião das pessoas em relação a uma marca ou um produto, como propõe o estudo realizado em (Li & Li, 2011).

Para atingir o objetivo de avaliar o sentimento presente em *tweets*, diversas estratégias já foram utilizadas, algumas fazendo uso das características do texto analisado. Como em (Pak & Paroubek, 2010), (Davidov et al, 2010), (Li & Li, 2011), (Zhang et al, 2011), (Jiang et al, 2011), por exemplo, onde *emoticons* – ícones que transmitem o estado emotivo da mensagem que acompanham – presentes nas mensagens foram utilizados como forma de reduzir o esforço de identificação da emoção presente. Além dos

emoticons, em (Davidov et al, 2010), as *hashtags* também foram utilizadas como forma de identificar o sentimento expresso no *tweet*. Já em (Hu et al, 2013), as relações entre os usuários e suas mensagens na rede do Twitter também foram consideradas para auxiliar a Análise de Sentimento. Essa estratégia baseou-se na premissa de que o sentimento é contagiante e, portanto, pessoas com relação social próxima podem expressar uma mesma emoção. Sendo assim, a relação de sentimento entre as mensagens foi utilizada como uma nova estratégia para lidar melhor com as características dos textos extraídos do Twitter.

Por fim, como um exemplo do uso de uma estratégia tradicional de Análise de Sentimento em *tweets*, um dos primeiros trabalhos cujo foco era a análise de emoções no Twitter buscou associar importantes eventos socioeconômicos e políticos às oscilações de sentimento identificadas nas mensagens (Bollen et al, 2011). Tal objetivo foi alcançado através da comparação de termos presentes nos textos com termos previamente associados à 6 categorias de humor, tornando possível a classificação dos tweets de acordo com cada um desses 6 sentimentos distintos.

2.5 Conclusão

Neste capítulo, os conceitos e as aplicações que envolvem o Twitter foram mostrados e foi possível identificar diversas iniciativas de estudo em relação à análise das opiniões expressas nas publicações realizadas através desta ferramenta. Dentre os pontos citados, destacamos a dificuldade em lidar com a escrita informal dos textos produzidos, o que se tornou um novo desafio para os que desejam explorar o sentimento presente no conteúdo de *microblogs* (Brew et al, 2011).

Através desta breve revisão da literatura, também ficou clara a baixa produção de material relacionado à análise de textos em português se comparado à língua inglesa, o que evidencia a importância deste trabalho. Em parte, isso se explica pela natureza informal do texto, agravado pelo menor número de recursos – dicionários, bases de teste etc. – disponíveis neste idioma. Sendo assim, este trabalho visa fornecer um novo recurso capaz de auxiliar na classificação destes textos, tornando possível também incorporar os termos informais encontrados ao processo de classificação da orientação das opiniões analisadas.

3 Análise de Sentimento

3.1 Introdução

Dentro da área de Recuperação da Informação, a tarefa de analisar "as opiniões, sentimentos, avaliações, apreciações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos" é conhecida por Análise de Sentimento (Liu, 2012). Esta atividade busca extrair sentimentos presentes em textos subjetivos e computacionalmente analisar e entender se aquela informação transmite uma opinião positiva, negativa ou neutra sobre um determinado assunto ou objeto de interesse (Missen et al, 2013). Por se tratar de um problema de Processamento de Linguagem Natural, a Análise de Sentimento engloba diversas atividades de análise sintática e semântica do texto para atingir seu objetivo, como veremos mais adiante.

Apesar de atualmente esta área de estudo ser conhecida principalmente por esta denominação, ao longo de sua evolução ela já recebeu diversos outros nomes que estavam intimamente relacionados às atividades específicas do processo de análise do sentimento e que, portanto, apesar de representarem igualmente uma mesma área de estudo, possuíam ligeiras diferenças em suas definições (Liu, 2012). Dentre eles, podemos destacar os nomes Mineração de Opinião, cujo foco é a extração de opinião presentes em textos, Análise de Subjetividade, que busca diferenciar textos subjetivos de textos objetivos presentes em um documento, e, por fim, Análise de Sentimento, que tem por objetivo classificar a polaridade de uma opinião (Pang & Lee, 2008). Os termos Análise de Sentimento e Mineração de Opinião surgiram na mesma época, em 2003, e acabaram por ser os mais adotados atualmente, apesar de Análise de Sentimento ainda ser o mais utilizado para representar esta área de estudo (Pang & Lee, 2008).

Neste capítulo, exploraremos esta área que atualmente é considerada uma das mais desafiadoras do campo de Recuperação da Informação (Missen et al, 2013). Para isso, na seção 3.2 definiremos o problema tratado pela Análise de Sentimento e apresentaremos os desafios da área, além de mostrarmos como eles já foram abordados por outros trabalhos. Na seção 3.3 trataremos da criação de léxicos para auxiliar a análise de sentimento. E na seção 3.4 faremos as conclusões finais deste capítulo.

3.2 Definição do problema e seus desafios

Uma das principais características do problema tratado pela Análise de Sentimento é o caráter subjetivo dos documentos por ela analisados. Diferente dos demais sistemas de Recuperação da Informação, o foco da área está no sentimento expresso no texto e não no fato a ele relacionado. Sendo assim, para que seja possível explorar os desafios e oportunidades oferecidos por esta área, primeiro é necessário entender o objeto central de toda a análise a ser realizada: a opinião.

Segundo (Liu, 2012), uma opinião possui a seguinte definição formal:

Definição (opinião): Uma opinião é uma quíntupla,

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l),$$

onde e_i é uma entidade, a_{ij} é um atributo de e_i, s_{ijkl} é um sentimento sobre a_{ij}, h_k é quem opina e t_l é o momento em que aquela opinião foi dada.

Esta representação demonstra que a opinião é uma emoção expressa por alguém em um determinado momento sobre algum assunto. Ela também expõe características essenciais que possuem extrema importância para a Análise de Sentimento e que são capazes de justificar não só o processo geral da análise de textos subjetivos, apresentado na Figura 1, mas também por que a ascensão desta área de estudo se deu somente a partir de 2001. Conforme dito anteriormente, esta área só ganhou destaque quando foi possível obter, entre outros fatores, um grande volume de dados subjetivos que pudessem ser analisados. Esta necessidade é facilmente justificada pelo fato de que cada opinião está relacionada à apenas um indivíduo. Sendo assim, para que os objetivos da análise de opiniões sobre um determinado tópico sejam alcançados, por exemplo, avaliar os prováveis resultados de eleições políticas, é preciso ter um conjunto de opiniões que sirva de amostra do que toda a população pensa em relação ao cenário político naquele momento. Portanto, a quantidade de opiniões a ser tratada deve ser grande o suficiente para ser capaz de representar essa população e pode ser um fator impeditivo do processo de Análise de Sentimento ainda atualmente.

Outra questão importante é o fato de que cada opinião está relacionada ao momento em que ela foi emitida. Esta característica justifica a importância de captar o sentimento do indivíduo no exato momento em que ele vivenciou o acontecimento que gerou aquela opinião. Dessa forma, a informação será captada antes que outros fatores a

influenciem e/ou diminuam a intensidade do sentimento gerado (Sloman et al., 2005) e o torne menos legítimo. Esta também é uma das razões pela qual ferramentas de *microblogging* têm se mostrado interessantes para a Análise de Sentimento, por permitirem que essa captação seja realizada em tempo real.

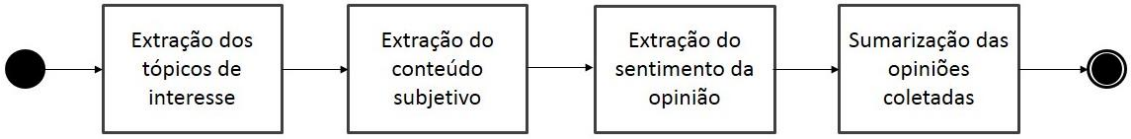

Figura 1 - Processo geral da Análise de Sentimento baseado nas atividades de criação de uma ferramenta de busca por opinião, conforme explicitado em (Pang & Lee, 2008).

Após definirmos formalmente a opinião, temos insumos suficientes para explorar os objetivos e os desafios das atividades que compõem o processo geral da Análise de Sentimento, que discutiremos a seguir.

3.2.1 Extração dos tópicos de interesse

A primeira atividade do processo tem por objetivo **identificar quais documentos fazem referência ao tópico de interesse da análise**, ou seja, a entidade à qual a opinião se refere. Esta atividade pode ser dividida em duas tarefas principais: a categorização e a extração da entidade (Liu, 2012).

O processo de categorização tem por objetivo levantar os termos que referenciam a entidade de interesse e os seus sinônimos, e fazer com que esse conjunto de palavras seja tratado como apenas um tópico (Nadeau & Sekine, 2007), como exemplificado na Tabela 1. A natureza dos textos tratados pela Análise de Sentimento faz desse processo um desafio, já que cada pessoa em seu texto pode designar o "apelido" que desejar para um determinado tópico, tornando muito difícil descobrir todos os termos sinônimos de um mesmo assunto ou objeto (Liu, 2012). Em (Li et al., 2010), esse problema foi tratado através da análise de alguns termos *seed* e das palavras próximas a eles no documento. Cada ocorrência de um desses termos gerou um vetor com a frequência das palavras consideradas na análise e um classificador Naïve-Bayes foi treinado para identificar novos candidatos ao conjunto de termos que referenciam a entidade. Este método, conhecido como *PU Learning*, conseguiu alcançar 72% de precisão.

Tabela 1 – Exemplos de associações entre termos do texto e o tópico de interesse

Tópico	Termos associados
Big Brother Brasil	big brother brasil
	big brother
	bbb
	bbb13

Já o processo de extração da entidade se assemelha à tarefa de Reconhecimento de Entidades Nomeadas (REN), uma subtarefa conhecida da Extração da Informação (Liu, 2012). Esse problema é anterior à Análise de Sentimento e já passou por diversas fases até alcançar seu estado atual (Sarawagi, 2008), onde estratégias baseadas em regras de extração e baseadas em modelos estatísticos trabalham juntas.

As técnicas baseadas nas regras de extração, em um primeiro momento, eram altamente dependente dos pesquisadores, já que tais regras precisavam ser definidas manualmente. Nesta estratégia manual, é necessário realizar uma análise do *corpus* e estabelecer regras capazes de extrair do texto a informação de interesse (Sarawagi, 2008). Um exemplo do uso desta técnica pode ser visto em (Silva et al, 2010), onde regras manualmente criadas são utilizadas em um sistema para extrair adjetivos encontrados no texto. No mesmo momento, já existia também a técnica baseada em aprendizado, onde um conjunto de textos rotulados manualmente servem de exemplo para os sistemas de aprendizagem de máquina aprenderem o padrão de texto que deve ser extraído dos demais documentos do *corpus* (Sarawagi, 2008).

Após essa fase, as técnicas baseadas em regras passaram a aprender automaticamente esses padrões de extração e, além disso, surgiram os modelos estatísticos (Sarawagi, 2008). Esses últimos, atualmente, baseiam-se em *Conditional Random Fields*, que são modelos estatísticos capazes de calcular a probabilidade de uma determinada sequência de rótulos ser atribuída à uma sequência de trechos do texto (Lafferty et al, 2001).

A junção das tarefas de categorização e extração dos tópicos de interesse permite a criação de um *corpus* com documentos que tratam do assunto de interesse da análise e que, portanto, deverão passar pelas próximas atividades do processo de Análise de Sentimento. Todo esse processo de extração de tópicos de interesse precisa ser realizado também para os atributos da entidade, quando os mesmos são considerados na análise realizada (Liu, 2012).

3.2.2 Extração do conteúdo subjetivo

A segunda atividade trata do problema de **separar os textos que contém opiniões daqueles que contém apenas fatos**. Inicialmente, este problema não era tratado de maneira automática e uma das primeiras iniciativas realizadas analisou, justamente, a

classificação manual de trechos de notícias em subjetivos ou objetivos (Bruce & Wiebe, 1999). O trabalho realizado permitiu compreender os padrões de concordância existentes entre as classificações realizadas e possibilitou, posteriormente, a utilização de uma estratégia de aprendizado supervisionado – quando instâncias são manualmente rotuladas para servirem de exemplo para os classificadores aprenderem o padrão de texto que se procura encontrar –, fazendo uso de um classificador Naïve-Bayes capaz de indicar a melhor classificação da subjetividade de um texto, quando o resultado obtido pela classificação manual não consegue atingir um consenso (Wiebe et al, 1999).

A partir do ano 2000, as estratégias para indicar a subjetividade de um documento passaram a analisar as propriedades das palavras existentes no texto para então identificar sentenças que possuem trechos opinativos. Essa técnica foi utilizada, por exemplo, em (Hatzivassiloglou & Wiebe, 2000), onde a orientação e a gradação de adjetivos – propriedade que indica se o adjetivo aceita ter seu significado intensificado ou diminuído por advérbios que os acompanham – foram utilizados como indicadores de subjetividade. Através de um modelo de regressão linear foi possível identificar o nível de gradação de cada adjetivo analisado e concluiu-se que a presença daqueles que permitem ter sua intensidade alterada é um bom indicativo de subjetividade da sentença. A orientação também se encaixa como um bom fator de subjetividade por estar sempre associada à sentenças que contém opinião. Esse estudo manteve o foco apenas em adjetivos, pois em (Bruce & Wiebe, 1999) foi constatado que a simples presença de adjetivos na frase indica 55,8% de probabilidade de aquela sentença apresentar conteúdo subjetivo. Sendo assim, partindo desta afirmação, o objetivo de (Hatzivassiloglou & Wiebe, 2000) era fazer uso de outras propriedades dessas palavras para melhorar os resultados obtidos por (Bruce & Wiebe, 1999), chegando a atingir acurácia de até 83% na predição de frases subjetivas.

Outras iniciativas de extração de conteúdo subjetivo buscaram diminuir a dependência de exemplos rotulados de maneira manual, como realizado em (Riloff et al, 2003) onde regras de extração de entidade foram utilizadas para identificar no texto, automaticamente, palavras subjetivas. O princípio é que palavras semanticamente similares são utilizadas em estruturas textuais parecidas e, portanto, a partir de um grupo de regras estabelecidas com base em palavras *seeds*, foi possível identificar outros termos considerados também subjetivos. Estes termos considerados subjetivos foram utilizados para classificar sentenças em subjetivas ou objetivas, com o auxílio de um classificador Naïve-Bayes.

Vale ressaltar que antes mesmo da popularização da área de Análise de Sentimento, a análise da subjetividade do texto era utilizada para auxiliar a Recuperação da Informação, conforme realizado em (Riloff et al, 2005). Através do trabalho realizado, foi possível observar que conteúdos subjetivos podem confundir as estratégias de extração da informação por apresentarem palavras que não representam seu significado literal naquele assunto, como casos em que metáforas são utilizadas, por exemplo. Sendo assim, conseguir separar textos subjetivos dos objetivos permitiu aumentar a precisão do sistema de recuperação em 3%, sem que a revocação do mesmo fosse prejudicada.

Já na área de Análise do Sentimento, esta atividade representa uma importante fase do processo e, portanto, novas estratégias foram testadas para tentar obter melhores resultados com a extração de textos subjetivos. Um exemplo foi o estudo realizado em (Pang & Lee, 2004) que fez uso de *minimum cuts* para distinguir frases subjetivas de frases objetivas. Essa técnica é baseada na análise de um grafo onde os nós representam as sentenças a serem classificadas e as arestas representam as associações entre essas frases, com peso indicativo da probabilidade de as duas pertencerem à uma mesma classe, ou as associações entre uma única sentença e as classes consideradas, nesse caso o peso de cada aresta indica a probabilidade da frase ser subjetiva ou objetiva. Tais probabilidades foram calculadas com base em um classificador Naïve-Bayes e o corte no grafo com menor custo, representado pela soma dos pesos das arestas, indica em quais pontos as sentenças se dividem entre subjetivas e objetivas.

Apesar de amplamente explorada, esta atividade ainda é vista como um desafio e isso se deve, principalmente, à falta de padronização dos textos analisados. Esta característica não permite estabelecer uma maneira única de identificar trechos subjetivos de forma independente de formato, estilo e nível de formalidade da escrita utilizada (Pang & Lee, 2008).

3.2.3 Extração do sentimento da opinião

A terceira atividade do processo de Análise de Sentimento tem como foco **avaliar se a opinião coletada expressa um sentimento positivo ou negativo e também qual é a intensidade desse sentimento**. A definição de opinião também nos auxilia nesta fase por deixar claro que cada opinião se refere à apenas uma entidade. É claro que às vezes o texto analisado faz referência e apresenta sentimentos sobre diferentes assuntos. Por isso, algumas iniciativas trabalham o texto previamente e passam a considerar sentenças

contendo apenas uma opinião como o documento objeto da análise. No final, esta é apenas uma diferença de estratégia, pois as mesmas técnicas se aplicam tanto às análises do documento como um todo quanto para trechos do texto original (Liu, 2012). Tais técnicas estão também baseadas em propriedades linguísticas dos textos analisados e ao longo desta seção exploraremos as mais frequentemente exploradas nos trabalhos da área.

As iniciativas da extração do sentimento, geralmente, utilizam aprendizado supervisionado, como feito em (Pang et al, 2002), uma das primeiras iniciativas de classificação de críticas de filmes em positivas ou negativas. O objetivo principal do estudo realizado foi avaliar se as técnicas de aprendizado de máquina utilizadas para a classificação de tópicos poderiam também ser utilizadas para a classificação de sentimentos ou se novos métodos precisariam ser desenvolvidos. Os testes foram feitos com classificadores Naïve-Bayes, de Máxima Entropia e SVM e, apesar de os resultados serem bons, foi verificado que a performance destes classificadores ficou abaixo da obtida para a tarefa de classificar tópicos, o que permitiu concluir que o nível de dificuldade da classificação de sentimentos é realmente maior.

Uma das iniciativas mais influentes da área, no entanto, fez uso de aprendizado não supervisionado para classificar filmes como recomendado ou não recomendado, através da análise de críticas dadas à eles. Em (Turney, 2002), a estratégia utilizada contou com um algoritmo chamado PMI-IR, capaz de avaliar a similaridade entre dois pares de palavras ou sentenças através do cálculo do Pointwise Mutual Information (PMI) entre esses textos. Este cálculo leva em consideração a frequência das palavras, uma das propriedades amplamente usadas nas técnicas de Análise de Sentimento. Esta estratégia também fez uso de uma outra propriedade linguística muito utilizada, o *Part-of-Speech* (POS) – função sintática exercida pela palavra no texto – para extrair bigramas compostos por um advérbio ou adjetivo e uma segunda palavra que conferisse contexto à sentença. Utilizando como parâmetro positivo de comparação da similaridade o termo “excelente”, e como parâmetro negativo o termo “ruim”, cada um dos bigramas foram analisados e aqueles que obtiveram uma similaridade maior com o termo “excelente” do que com o termo “ruim” foram considerados positivos. Os demais foram considerados negativos e as críticas foram classificadas com base na média das polaridades dos bigramas existentes em cada uma delas.

Em 2003, o PMI-IR foi estendido e utilizado novamente em (Turney & Littman, 2003). Neste novo estudo, os autores utilizaram não apenas uma palavra, mas um conjunto de 7 palavras positivas e 7 negativas como parâmetro de comparação. Tais palavras foram selecionadas por sua falta de dependência do domínio, ou seja, por manterem sua orientação original independente dos termos que as acompanham. Tal alteração na execução do algoritmo permitiu aumentar a acurácia obtida anteriormente em até 11%, chegando a atingir 95%. Esta pesquisa realizada por (Turney & Littman, 2003) serviu como base para esse trabalho e, portanto, será explorada detalhadamente mais adiante, no capítulo 4.

Diversas iniciativas buscam explorar as palavras presentes nos textos e, para isso, contam com o apoio de dicionários. Como veremos a seguir, as estratégias para lidar com esses dados mudam, porém é indiscutível a importância desse recurso para as iniciativas de Análise de Sentimento. Tal valor foi constatado em (Kamps et al, 2004), onde as relações de sinonímia e antonímia presentes no WordNet – principal dicionário disponível para pesquisas na língua inglesa – foram utilizadas para calcular a orientação semântica de adjetivos. O método calculava a polaridade do termo avaliado através da fórmula $EVA(w) = d(w, ruim) – (d(w,bom)/d(bom,ruim))$, onde d é o caminho mais curto entre as duas palavras no grafo criado a partir do WordNet. Palavras com $EVA(w) > 0$ eram consideradas positivas e as com $EVA(w) < 0$, negativas. O valor absoluto obtido ainda permitia designar a intensidade do sentimento presente naquele termo.

O trabalho realizado em (Kim & Hovy, 2004) faz uso de mais uma propriedade linguística do texto. Para classificar as sentenças analisadas ele gera um dicionário de polaridades com base no WordNet e, ao verificar a presença de palavras já conhecidas no texto, a existência de palavras de negação é considerada. Sendo assim, ao encontrar palavras como não ou nunca antecedendo um termo positivo, por exemplo, esse termo será considerado como negativo no cálculo final da polaridade da sentença. A criação do dicionário é feita com base no cálculo da probabilidade de uma determinada palavra pertencer às classes de palavras previamente rotuladas como positivas ou negativas.

Outro trabalho realizado com apoio no WordNet foi o de (Goldbole et al, 2007), cujo objetivo era o de classificar cada entidade encontrada em textos de blogs e notícias como positiva ou negativa. Assim como em (Kamps et al, 2004) e (Kim & Hovy, 2004),

esta iniciativa conta a criação prévia de um léxico de apoio. Tal estratégia serviu como base para este trabalho e será detalhada mais adiante, no capítulo 4.

Observando a literatura, fica claro que este é o grande foco das pesquisas da área, sendo também o objetivo principal deste trabalho. Também é possível observar que apesar das diversas iniciativas e bons resultados obtidos, esta ainda é uma atividade em evolução e em busca de melhores métodos que possibilitem maior acurácia na classificação do sentimento da opinião.

3.2.4 Sumarização das opiniões coletadas

Nesta última atividade, o foco está em **como agrupar e representar os resultados** obtidos através das análises realizadas anteriormente (Pang & Lee, 2008). Esta atividade é de extrema importância, pois o seu resultado gera o produto final desejado por aqueles que utilizam a Análise de Sentimento para auxiliar o seu processo de tomada de decisão (Hu & Liu, 2004).

A sumarização tradicional de textos pode ser realizada de duas maneiras: baseada em apenas um documento ou baseada em um conjunto de documentos (Pang & Lee, 2008). Apesar de já ter sido utilizada com textos opinativos (Pang & Lee, 2008), a primeira estratégia não é a que melhor se aplica à Análise de Sentimento, e a justificativa para isto está na própria definição de opinião, discutida anteriormente. Conforme falamos, para obter o sentimento de um grupo de pessoas em relação à algum assunto, precisamos analisar um conjunto de opiniões e não apenas uma opinião isolada. Sendo assim, a estratégia de sumarização baseada em um conjunto de documentos se aplica melhor quando estamos tratando textos subjetivos (Liu, 2012).

O fato de estarmos lidando com opiniões ainda exige uma diferenciação no que diz respeito às técnicas de sumarização utilizadas (Liu, 2012). Em métodos tradicionais, o resultado da sumarização de apenas um documento costuma ser a criação de um novo texto, mais curto, contendo as principais informações que foram extraídas do original. No caso do método que faz uso de múltiplos documentos, o resultado costuma trazer a diferença entre eles, descartando trechos considerados similares. Para a opinião, seguindo sua definição, precisamos tratar as entidades e os atributos existentes, além do percentual de opiniões positivas e negativas relacionadas à cada um deles (Liu, 2012).

Seguindo este direcionamento, o trabalho realizado em (Hu & Liu, 2004) buscou sumarizar críticas de produtos de forma a obter um resultado que indicasse a quantidade de críticas positivas e negativas que cada atributo de cada produto avaliado pelos consumidores receberam. Para isso, primeiro foi necessário extrair os produtos e seus atributos citados nas críticas, seguindo técnicas de extração similares às já discutidas anteriormente. O segundo passo foi relacionar as opiniões às entidades extraídas e identificar a orientação dessas opiniões. Por fim, a sumarização dos resultados foi realizada de forma que cada atributo encontrado foi apresentado estando relacionado não somente ao produto como também às opiniões coletadas. Conforme podemos verificar na Figura 2, este resumo estruturado nos permite visualizar, rapidamente, se o sentimento em relação àquele atributo é positivo ou negativo, atingindo o objetivo final da etapa de sumarização.

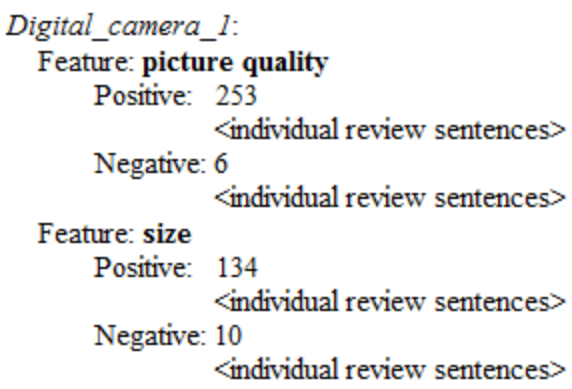

Figura 2 – Adaptação do exemplo de sumarização de opiniões proposto por (Hu & Liu, 2004).

Outras formas de sumarizar opiniões também já foram estudadas, como, por exemplo, a sumarização de pares de opiniões contrastantes. Em (Kim & Zhai, 2009), os autores exploraram essa forma de sumarização, onde o objetivo é encontrar sentenças sobre um mesmo atributo de uma entidade, porém com sentimentos opostos. Para isso, as sentenças precisam ser clusterizadas em positivas e negativas e, em um segundo momento, a similaridade entre elas precisa ser calculada. Em (Kim & Zhai, 2009), duas funções de similaridade foram utilizadas para avaliar as sentenças no nível das palavras presentes, uma verificando a existência de um mesmo termo nas duas sentenças e a outra verificando se os termos comparados eram similares semanticamente. Este método de agrupamento e representação das opiniões coletadas gera um resumo similar ao exemplificado na Tabela 2.

Tabela 2 – Exemplo de sumarização por contraste adaptado de (Kim & Zhai, 2009).

No	Positive	Negative
1	oh... and file transfers are fast & easy.	you need the software to actually transfer files
2	I noticed that the micro adjustment knob and collet are well made and work well too.	the adjustment knob seemed ok, but when lowering the router, i have to practically pull it down while turning the knob.
3	The navigation is nice enough, but scrolling and searching through thousand of tracks, hundreds of albums or artists, or even dozens of genres is not conducive to save driving	difficult navigation – i wo n't necessarily say " difficult ," i do n't enjoy the scrollwheel to navigate
4	I imagine if i left my player untouched (no backlight) it could play for considerably more than 12 hours at a low volume level	there are 2 things that need fixing first is the battery life. it will run for 6 hrs without problems with medium usage of the buttons

É importante ressaltar que a atividade de sumarização das opiniões coletadas é altamente dependente dos resultados e das evoluções das atividades que a antecedem (Liu, 2012), já que o formato e a qualidade dos dados obtidos por essas fases influencia diretamente a maneira como essas informações serão apresentadas. Além disso, as sumarizações realizadas não precisam estar focadas apenas em representar os resultados através de textos. Essa representação também pode ser realizada por meio de gráficos, por exemplo, fazendo uso das estatísticas dos resultados (Pang & Lee, 2008).

3.3 Criação de dicionários para apoio à Análise de Sentimento

Ao discutirmos as atividades do processo de Análise de Sentimento ficou clara a enorme influência dos dicionários para a área. Esses recursos costumam oferecer informações sobre as relações semânticas e/ou as polaridades das palavras e podem ser utilizados em diversas técnicas auxiliando, por exemplo, tarefas que buscam identificar expressões subjetivas no texto. A grande importância do dicionário está em permitir melhorias nos resultados sem grande esforço adicional, pois ele possui documentadas informações sobre palavras já conhecidas, podendo esses dados estarem atrelados a um determinado tópico ou não (Souza et al, 2011).

Na língua inglesa, o principal dicionário existente é o WordNet. Atualmente em sua terceira versão, ele possui cerca de 155.280 termos, entre eles substantivos, verbos, adjetivos e advérbios, agrupados de acordo com seus significados semânticos. Sendo assim, o WordNet fornece para os pesquisadores da área informação sobre as relações de sinonímia e antonímia entre um enorme conjunto de palavras e, como pudemos observar na seção anterior, é amplamente utilizado em diversas técnicas de Análise de Sentimento (WordNet, 2013).

Em 2006, o trabalho realizado em (Esuli & Sebastiani, 2006) fez uso do WordNet para criar um novo dicionário contendo as polaridades das palavras, o SentiWordNet. A técnica utilizada é conhecida como "estratégia baseada em dicionário", ou seja, as informações de um dicionário já existente são utilizadas para auxiliar na criação de um novo léxico. No caso de (Esuli & Sebastiani, 2006), eles partiram de um pequeno conjunto de palavras *seed* positivas e negativas e utilizaram o algoritmo publicado em (Esuli & Sebastiani, 2005) para calcular as novas polaridades. Nesse algoritmo, as relações de sinonímia e antonímia foram usadas para enriquecer o conjunto de palavras *seed* e, posteriormente, as definições textuais dos significados das palavras presentes no WordNet foram consideradas para calcular a similaridade – através de Similaridade por Cosseno – entre novos termos avaliados e os termos *seed*. Por fim, as novas palavras encontradas e classificadas foram adicionadas ao SentiWordNet.

Dicionários gerados com base em outros pré-existentes possuem a característica de serem independentes de domínio. Em Análise de Sentimento esta característica pode ser considerada desvantajosa, pois a orientação de uma palavra pode ser alterada dependendo do domínio do texto em que ela se encontra. Uma alternativa para este problema é utilizar a "estratégia baseada no *corpus*" para criar o novo dicionário e assim considerar o domínio ao longo do processo (Liu, 2012). Esta técnica foi utilizada, por exemplo, em (Hatzivassiloglou & McKeown, 1997), para gerar um dicionário de polaridades de adjetivos extraídos do *corpus* analisado. Para isso, um conjunto de adjetivos que possuíam maior frequência no *corpus* foram utilizados para, através da análise de conjunções, extrair novos adjetivos do texto. Esses mesmos adjetivos foram também manualmente classificados e os novos termos que estivessem conectados por "e" foram considerados de mesma classe, e aqueles ligados por "mas" de classe oposta.

Apesar de na língua inglesa já existirem alguns léxicos bastante ricos disponíveis, o mesmo não acontece para o português. Uma iniciativa bastante interessante foi a criação do TeP 1.0 e, posteriormente, a evolução para o TeP 2.0. Esse dicionário foi gerado com estrutura similar ao WordNet e atualmente possui cerca de 44.670 termos, com suas relações de sinonímia e antonímia. Um ponto importante do TeP 2.0 é o fato de ele ser todo em português do Brasil e, por isso, ele será utilizado como um dicionário base deste trabalho (Maziero et al, 2008).

Em (Souza et al, 2011), o TeP 2.0 foi utilizado como base para a criação de um novo dicionário, juntamente com outros léxicos. O objetivo dos autores era gerar um dicionário de polaridades, em português de Portugal, somente com adjetivos e expressões idiomáticas capazes de classificar pessoas. Para isso, um grafo de palavras *seed* foi gerado com base nas relações existentes nos léxicos *seed* e algumas polaridades iniciais foram atribuídas manualmente. Além disso, os autores ainda utilizaram regras de extração para obter novos adjetivos classificadores de características humanas e agregaram ao novo dicionário. A partir daí, a similaridade entre os termos foi calculada através da distância de menor caminho de Dijkstra entre o termo avaliado e um termo que representava cada uma das classes de orientação. Esse modelo conseguiu atingir a acurácia de 67% para termos positivos, 45% para termos neutros e 82% para termos negativos e também servirá como base para o método utilizado neste trabalho.

Apesar da inquestionável importância dos léxicos para as iniciativas da área de Análise de Sentimento, infelizmente não é de nosso conhecimento a existência de um dicionário com polaridades em português do Brasil. Em português de Portugal existe ainda a iniciativa de (Santos et al, 2011), porém tal recurso não se adequa totalmente à textos em português do Brasil devido à diferença na escrita de determinadas palavras. Por esse motivo, este trabalho tem por objetivo gerar um novo dicionário para que os trabalhos realizados nesse idioma possam usufruir de mais esse recurso linguístico em suas análises.

3.4 Conclusão

Neste capítulo, os conceitos, os desafios e as iniciativas já realizadas em Análise de Sentimento foram explorados. É possível perceber que esta é uma área ainda em evolução e que, a cada dia, novas estratégias e novos recursos linguísticos são utilizados com o objetivo de aprimorar os resultados obtidos através da classificação automática de opiniões escritas por humanos e atingir novos patamares de satisfação dos pesquisadores com os resultados obtidos.

Vale ressaltar que recentemente este campo de estudo deixou de receber interesse apenas da comunidade acadêmica, mas passou a atrair a atenção também de empresas que buscam entender o que as pessoas pensam sobre seus produtos para definir suas próximas ações comerciais. Estas necessidades unidas aos desafios ainda existentes para os pesquisadores da área prometem manter esta área de estudo viva (Liu, 2012), além de

servir de justificativa suficiente para a criação de novos recursos linguísticos com foco em português do Brasil.

4 Criação de dicionário em português do Brasil

No capítulo anterior, alguns trabalhos da área de Análise de Sentimento foram explorados. Tal análise permitiu constatar não só a importância dos dicionários para as estratégias propostas em soluções anteriores, como também a necessidade da criação deste recurso para o idioma português do Brasil. Por este motivo, este trabalho apresenta a abordagem utilizada para gerar sete novos dicionários, utilizando diferentes métodos. Neste capítulo, tais métodos serão explorados detalhadamente de forma que seja possível entender como foi o processo de criação de cada um desses dicionários e, também, justificar o uso de cada estratégia.

Vale ressaltar que este trabalho teve como base o estudo realizado em (Silva et al, 2010), porém se diferencia desse ao possibilitar a criação de dicionários em diferentes domínios, ao realizar adaptações às técnicas previamente utilizadas por (Silva et al, 2010) para gerar um léxico com foco em pessoas. Além disso, este trabalho ainda permite realizar uma comparação de diferentes métodos de cálculo de polaridade e, ao final, permite identificar aquele que obtém melhor resultado na classificação de palavras em positivas, negativas e neutras. Por fim, como dito anteriormente, este trabalho mantém o foco no idioma português do Brasil, diferentemente de (Silva et al, 2010), cujo foco era o português de Portugal.

Este capítulo apresenta as técnicas de criação do dicionário *seed* e dos métodos de cálculo de polaridade utilizados. Para isto, na seção 4.1 deste capítulo apresentaremos o processo de criação de dicionários como um todo; na seção 4.2, discutiremos os dicionários utilizados para a criação do dicionário *seed*; na seção 4.3, exploraremos a técnica de inclusão de termos do *corpus*; na seção 4.4, definiremos os métodos utilizados para calcular a orientação dos termos do dicionário; e, por fim, a seção 4.5 apresenta as considerações finais deste capítulo.

4.1 O processo de criação do dicionário

Antes de abordamos cada atividade do processo de criação do dicionário de maneira mais detalhada, na Figura 3 apresentamos uma visão geral de todos os passos necessários para a geração deste recurso. São ao todo 6 passos que, no final, nos permitem gerar um dicionário de polaridades adaptado ao domínio que se deseja explorar.

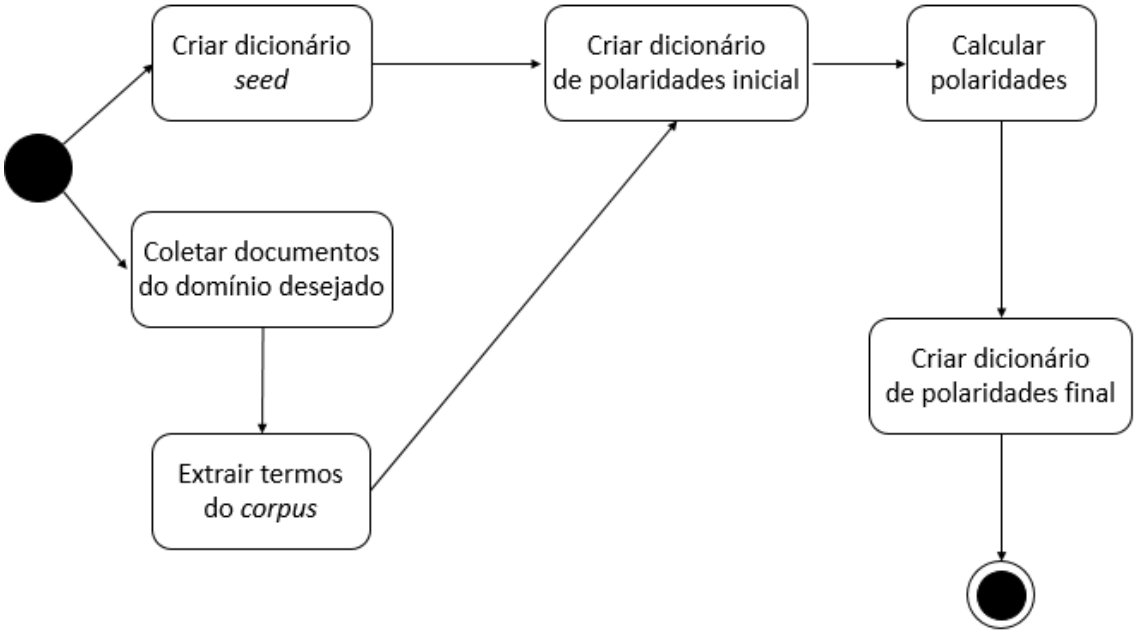

Figura 3 – Processo de criação do dicionário de polaridades.

Como podemos observar na Figura 3, as duas primeiras atividades, que podem ocorrer em paralelo, dizem respeito à geração de um dicionário *seed* e à coleta de documentos relacionados ao tópico que se deseja explorar. A primeira atividade é necessária porque a descoberta das demais polaridades do dicionário precisam partir de um ponto inicial. É preciso ter algum conhecimento sobre pelo menos parte dos termos que compõem o dicionário para que eles sirvam de origem para os demais termos ainda não classificados, e também para aqueles que serão extraídos do *corpus* posteriormente. Sendo assim, esta etapa pode ser realizada de forma manual, coletando e classificando termos, ou pode ser feita utilizando recursos já existentes para o idioma que se deseja tratar. Já a segunda atividade, a coleta de documentos, gerará um importante *corpus* que fornecerá dados para que novos termos, específicos do domínio analisado, sejam encontrados, o que é realizado na etapa de extração de termos do *corpus*.

Após essas 3 atividades, é possível construir o que chamamos de dicionário de polaridades inicial. Este dicionário já contém todos os seus termos, porém nem todos estão classificados. Na realidade, neste momento, só possuem orientação aqueles que foram manualmente classificados na etapa de criação do dicionário *seed*, ou que tiveram essa informação extraída de algum outro léxico previamente existente. A partir deste ponto, todas as relações e informações obtidas sobre os termos até o momento serão utilizadas para realizar o cálculo das polaridades dos termos presentes no dicionário, a próxima atividade do processo.

Neste trabalho, foram utilizadas 7 maneiras de calcular essa orientação e, por isso, falamos que foram gerados 7 dicionários de polaridades distintos, já que cada estratégia utilizada obteve o seu resultado. Portanto, fez-se necessária a última etapa, a de criação do dicionário de polaridades final, onde todos os resultados obtidos são analisados e aqueles com melhor desempenho são adotados como a polaridade do dicionário final gerado.

Vale ressaltar que, para este trabalho, conseguir calcular a orientação das demais palavras presentes no grafo é de extrema importância, já que as polaridades obtidas através do SentiLex contemplam apenas 11% dos termos presentes no TeP 2.0, e é ele que possui termos unicamente em português do Brasil.

Nas próximas seções, todas essas atividades serão apresentadas com maior detalhe. No entanto, já é possível perceber que o processo aqui apresentado permite gerar novos dicionários em qualquer idioma ou com base em qualquer domínio, permitindo que a estratégia geral aqui explorada seja utilizada em outras oportunidades, possibilitando a criação de léxicos voltados para o interesse da pesquisa em questão.

4.2 Criação do dicionário *seed*

Seguindo uma das técnicas de geração de novos dicionários e também a estratégia utilizada em (Silva et al, 2010), o primeiro passo da abordagem utilizada neste trabalho consta da construção de um dicionário *seed* criado com base em léxicos já existentes, sendo eles o TeP 2.0 e o SentiLex. O primeiro, composto apenas por termos do português do Brasil, é unido ao segundo para extrair desse as polaridades iniciais de parte das palavras. Tais polaridades servirão como ponto de partida para os algoritmos de cálculo de polaridade, que detalharemos mais à frente.

Para entender melhor a estrutura desse dicionário *seed*, vamos, primeiramente, analisar o processo de criação de cada um dos léxicos de origem.

4.2.1 TeP 2.0: um *thesaurus* eletrônico para o Português do Brasil

O TeP 2.0 é a segunda versão do TeP, um *thesaurus* eletrônico de sinônimos e antônimos, criado com o objetivo de oferecer aos usuários da língua portuguesa do Brasil um dicionário, disponível na Internet, onde fosse possível buscar opções de palavras sinônimas e antônimas para o uso na produção de textos (Dias-da-Silva & Moraes, 2003).

Composto por substantivos, adjetivos, advérbios e verbos, ele foi criado manualmente de acordo com a estrutura do WordNet e possui na sua base informações provenientes do WordNet.Br, uma versão do WordNet para o Português do Brasil que possui, até o momento, apenas verbos (Dias-da-Silva, 2010).

Por seguir a formatação do WordNet, o TeP 2.0 apresenta seus termos agrupados em *synsets* (*synonym sets*). Cada *synset* reúne palavras que representam um mesmo conceito e "que podem ser intercambiáveis em um determinado contexto" (Maziero et al, 2008), além de serem de mesma classe gramatical. Para gerar os 19.888 synsets que atualmente compõem o TeP 2.0 (Maziero et al, 2008), cinco grandes dicionários da língua portuguesa do Brasil foram analisados de forma que os significados de cada vocábulo foram utilizados para extrair os termos sinônimos candidatos, como podemos visualizar nas Figuras 4 e 5 abaixo (Dias-da-Silva & Moraes, 2003).

lembrar
v. 1. Tr. dir. Trazer à memória; recordar. 2. Tr. ind. Vir à idéia, tornar-se recordado. 3. Pron. Recordar-se, ter lembrança de. 4. Tr. dir. Fazer vir à memória por analogia ou semelhança. 5. Tr. dir. Advertir, notar. 6. Tr. dir. Sugerir. 7. Tr. dir. Recomendar.

Figura 4 – Exemplo de verbete analisado para extrair conjuntos de sinônimos, adaptado de (Dias-da-Silva & Moraes, 2003).

{lembrar, recordar}
{lembrar, advertir, notar}
{lembrar, sugerir}
{lembrar, recomendar}
{lembrar-se, recordar-se}

Figura 5 – Exemplo de conjuntos de sinônimos extraídos do verbete da Figura 3, de acordo com os diferentes significados identificados, adaptado de (Dias-da-Silva & Moraes, 2003).

Esse processo permitiu a criação de conjuntos de sinônimos que, após serem validados, eram adicionados ao dicionário. Essa validação era feita utilizando o mesmo procedimento para cada um dos termos sinônimos candidatos. Caso os mesmos conjuntos de sinônimos fossem encontrados ao analisar a palavra candidata, tal conjunto era considerado legítimo. Essa análise foi realizada com todos os termos encontrados nos dicionários base analisados (Dias-da-Silva & Moraes, 2003).

Foi também essa estratégia que identificou as relações de antonímia entre os conjuntos de sinônimos previamente estabelecidos. Ao analisar o significado de dois termos, é possível verificar paráfrases com sentidos opostos, como podemos visualizar ao comparar os dois verbetes da Figura 6. Essa informação permite identificar, de acordo

com o exemplo, que esquecer é antônimo de lembrar. É possível também concluir que, se considerarmos um mesmo significado, todos os sinônimos de esquecer são antônimos de todos sinônimos de lembrar, justificando assim a razão de a relação de antonímia ser estabelecida entre *synsets* e não entre vocábulos individuais (Dias-da-Silva & Moraes, 2003).

lembrar
v. 1. Tr. dir. Trazer à memória; recordar. 2. Tr. ind. Vir à idéia, tornar-se recordado. 3. Pron. Recordar-se, ter lembrança de. 4. Tr. dir. Fazer vir à memória por analogia ou semelhança. 5. Tr. dir. Advertir, notar. 6. Tr. dir. Sugerir. 7. Tr. dir. Recomendar.

esquecer
v. 1. Tr. dir. Deixar sair da memória; perder a memória de; tirar da lembrança; olvidar. 2. Pron. Perder a lembrança ou a memória; olvidar-se. 3. Tr. dir. Não fazer caso de, pôr em esquecimento. 4. Tr. ind. e intr. Escapar da memória, ficar em esquecimento: Esqueceu-lhe o final do discurso. Seu prestígio foi momentâneo, passou e esqueceu. 5. Tr. dir. Descurar-se de: Não esquecia as suas tarefas. 6. Pron. Perder a ciência ou a habilidade adquiridas: Já me esqueci do latim. 7. Pron. Descuidar-se: Meu secretário esqueceu-se de tudo. 8. Intr. Ficar dormente ou tolhido, perder a sensibilidade: Naquela má posição a perna esqueceu.

Figura 6 – Exemplo de dois verbetes antônimos deixando explícitas as paráfrases que indicam tal relação semântica, adaptado de (Dias-da-Silva & Moraes, 2003).

Atualmente, o TeP 2.0 é disponibilizado através de uma interface *web* que permite ao usuário realizar consultas e visualizar os *synsets* existentes, além dos exemplos de uso extraídos do WordNet.Br, quando presentes. O TeP 2.0 pode ainda ser incorporado a qualquer sistema que busque um recurso capaz de fornecer relações de antonímia e sinonímia entre as palavras. Sua base é disponibilizada em um arquivo de formato texto e segue a formatação apresentada na Figura 7 abaixo. Nela podemos observar que os *synsets* possuem apenas termos de mesma classe gramatical, conforme citado anteriormente, e que os sinônimos de cada grupo são separados por vírgula. A informação do *synset* antônimo, quando existe, é apresentada entre símbolos de maior e menor e mostra o número do *synset* que realiza a função de antônimo do *synset* em questão.

1. [Verbo] {exagerar, exceder, quinta-essenciar, rebuscar, refinar, requintar}
2. [Verbo] {ababalhar, babar, escumar, espumar}
3. [Verbo] {ababalhar, sujar}
4. [Verbo] {aclarar, alumiar, clarear, iluminar}
5. [Verbo] {enegrecer, escurecer} <4>
6. [Verbo] {limpar, purgar, purificar}
7. [Verbo] {impurificar, sujar} <6>
8. [Verbo] {apurar, clarificar, limpar, purgar, purificar}
9. [Verbo] {purgar} <7>

Figura 7 – Trecho extraído da base do TeP 2.0 para demonstrar a formatação de apresentação dos *synsets*.

4.2.2 SentiLex: um dicionário em Português de Portugal, com foco em pessoas

O SentiLex é um léxico composto por adjetivos capazes de classificar e/ou modificar substantivos que representam características humanas, como por exemplo, profissões, títulos e pronomes pessoais. Um diferencial deste dicionário é a presença de adjetivos extraídos do texto através do uso de regras manualmente criadas com foco no tópico tratado pelo dicionário (Silva et al, 2010). Conforme dito anteriormente, o processo de criação do SentiLex serviu como base para este trabalho, sendo adaptado quando necessário para melhor atender aos objetivos aqui estabelecidos.

Tal processo de criação conta com dois principais insumos: um léxico de adjetivos com suas categorias semânticas (humano ou não humano) e polaridades (positivo, negativo e neutro) parcialmente classificadas de forma manual; e dicionários de sinônimos e antônimos, dentre eles o TeP 2.0. Além destes, também foram utilizados dicionários de nomes, sobrenomes e de profissões, que serviram para criar as regras de extração de adjetivos, que veremos detalhadamente mais adiante, na próxima seção.

As regras estabelecidas permitiram extrair, de um corpus de cerca de dez milhões de documentos escritos em português, mais de 8.500 adjetivos. Após a fase de extração, foi necessário validar se tais adjetivos realmente estariam relacionados a substantivos com categoria semântica humana. Para isso, foram utilizadas quatro métricas: a quantidade de vezes que o adjetivo identificado aparecia no corpus; o número total de regras com as quais ele casou; a frequência de cada regra com a qual ele casou; e quantas vezes ele casou com alguma regra. Essas validações permitiram separar os adjetivos altamente conectados à termos humanos dos que provavelmente eram não humanos. Os primeiros foram adicionados ao dicionário.

Esse processo de união dos léxicos de sinônimos com os adjetivos extraídos do texto gerou o SentiLex, ainda parcialmente polarizado de forma manual. Para calcular a polaridade dos demais termos, as relações entre as palavras foram consideradas e, através do cálculo da distância de caminho mais curto de Dijkstra, as novas polaridades foram calculadas. Este cálculo levou em consideração as palavras antes manualmente classificadas em positivo, negativo e neutro e, ao avaliar um adjetivo, aquela classe que estivesse mais próxima a ele seria a classificação designada.

Pelo fato de o SentiLex ter sido criado com base no TeP 2.0, é possível realizar a união desses dois léxicos e trazer para os termos do TeP 2.0 as polaridades estabelecidas no SentiLex. Essa união nos permite obter um dicionário *seed* em português do Brasil, parcialmente polarizado e sem influência do domínio analisado. No entanto, conforme já observado em (Qiu et al, 2012), introduzir termos do domínio ao dicionário é altamente desejado, pois nos permite ter informações sobre a forma de utilização de termos já conhecidos dentro dos textos analisados, além de tornar possível incorporar ao dicionário novas palavras, específicas do assunto tratado. Para isso, a segunda etapa da abordagem, onde a inclusão de termos do domínio pudessem ser incorporados ao dicionário, se tornou necessária, conforme aconteceu em (Silva et al, 2010).

4.3 Inclusão de novos termos oriundos do *corpus*

Esta etapa é de grande importância para a criação de qualquer dicionário que tenha a intenção de refletir características específicas do domínio tratado. No caso deste trabalho, esta etapa é ainda mais importante devido à natureza dos textos selecionados. Por se tratarem de textos oriundos do Twitter, diversos termos utilizados não se encontravam no dicionário inicial, como gírias e outras construções específicas dos usuários da plataforma. Sendo assim, esta etapa permitiu superar esta deficiência, além de facilitar o tratamento de palavras escritas de maneira errada.

A técnica utilizada para a extração desses termos foi a mesma aplicada em (Silva et al, 2010), onde regras sintáticas foram criadas para identificar adjetivos no texto e incorporá-los ao dicionário. Na Tabela 3, podemos observar essas regras e perceber que grande parte delas eram dependentes de substantivos humanos, conforme mencionado na seção anterior. Tais substantivos estão identificados pelas seguintes siglas:

- HREF - substantivos humanos genéricos, como pessoa e indivíduo;

- ERGO – nomes de profissões ou cargos, como primeiro-ministro e professor;
- IREF – pronomes pessoais de terceira pessoa no singular;
- N – primeiro nome;
- N S – combinação de primeiro nome e sobrenome;
- S S – combinação de dois sobrenomes seguidos um do outro;

Além disso, ainda são considerados verbos copulativos, como ser e estar, representados através da sigla COP, e advérbios e intensificadores, identificados pela sigla MODIF.

Tabela 3 – Tabela de regras utilizadas no SentiLex, adaptado de (Silva et al, 2010).

ID	Padrão	ID	Padrão	ID	Padrão
301	N COP ADJ	**401**	N S COP ADJ	**501**	N S COP MODIF ADJ
302	IHREF COP ADJ	**402**	S S COP ADJ	**502**	S S COP MODIF ADJ
303	HREF COP ADJ	**403**	N COP MODIF ADJ	**503**	N S é um|uma ADJ
304	ERGO COP ADJ	**404**	N é um|uma ADJ	**504**	S S é um|uma ADJ
305	tu és|estás ADJ	**405**	o|a ADJ do|da N	**505**	o|a ADJ do|da S S
306	um|uma HREF ADJ	**406**	o|a ADJ do|da ERGO		
307	um|uma ERGO ADJ	**407**	tu és|estás MODIF ADJ		
308	um|uma ADJ HREF	**408**	HREF COP MODIF ADJ		
309	um|uma ADJ ERGO	**409**	ERGO COP MODIF ADJ		
		410	IREF COP MODIF ADJ		
		411	IREF é um|uma ADJ		
		412	HREF é um|uma ADJ		
		413	ERGO é um|uma ADJ		
		414	um|uma HREF MODIF ADJ		
		415	um|uma ERGO MODIF ADJ		

Através da tabela acima, também podemos perceber que as regras foram aplicadas a três grupos distintos de n-grama e, no SentiLex, as regras relacionadas aos trigramas produziram os melhores resultados, extraindo 68% dos 8.579 novos adjetivos, distintos entre si, coletados do *corpus* analisado.

No caso deste trabalho, as regras para trigramas também se destacaram, sendo responsável pela extração de 81% dos 7.646 novos adjetivos encontrados nos *tweets* analisados. No entanto, para se adequar às nossas necessidades, as regras do SentiLex precisaram ser adaptadas para se tornarem mais genéricas e conseguirem extrair os adjetivos presentes nos textos analisados de forma menos dependente do domínio

humano. Além disso, a partir da observação dos documentos, algumas novas regras foram propostas. Tais regras são apresentadas na Tabela 4 abaixo, junto com as demais.

Tabela 4 – Tabela de regras utilizadas neste trabalho.

ID	Padrão	ID	Padrão
301_ADP	WORD COP ADJ	**403_ADP**	WORD COP ADV ADJ
302	IREF COP ADJ	**404_ADP**	WORD COP um|uma ADJ
306_ADP	um|uma WORD ADJ	**408_ADP**	IREF COP ADV ADJ
308_ADP	um|uma ADJ WORD	**414_ADP**	um|uma WORD ADV ADJ
31	DET N ADJ	**41**	V N ADJ N
32	V N ADJ		
33	ADJ DET N		

As regras propostas neste trabalho podem ser identificadas pela numeração, que diferem da utilizada na identificação das regras do SentiLex, e pelo fato de basearem-se apenas na função gramatical exercida pelas palavras. As sequências que definiram esses padrões de extração foram observadas repetidamente nos textos analisados e, portanto, foram adicionadas ao conjunto de regras utilizadas para extrair novos adjetivos do texto. Vale ressaltar que o conjunto final de padrões utilizados neste trabalho não contou com todas as regras estabelecidas no SentiLex. Isso se deve ao fato de, ao generalizar algumas dessas regras, elas acabarem por englobar as demais, como no caso da 301_ADP, que engloba as originais 301, 303 e 304. Neste caso específico, achamos interessante manter a 302 conforme a definição original para que a precisão deste padrão fosse maior, já que a ocorrência do mesmo tende a ser maior também. Por fim, resolvemos não utilizar as regras de pentagramas por observarmos que sua influência no resultado final não era significativa, possivelmente devido ao estilo de escrita do texto tratado.

Ao final desta etapa obtivemos uma segunda versão do dicionário inicial, agora com influência do domínio analisado. Para completar a geração do dicionário de polaridades em português do Brasil, era necessário ainda calcular as polaridades para aqueles termos ainda não classificados.

4.4 Cálculo das polaridades do dicionário gerado

Ao realizarmos a revisão da literatura, identificamos trabalhos que obtiveram resultados mais interessantes que os obtidos por (Silva et al, 2010), onde a acurácia do sistema alcançou 67%. Sendo assim, resolvemos adaptar essas estratégias para avaliar a sua eficácia ao serem aplicadas ao dicionário gerado pelos passos anteriores. Além disso, decidimos também adicionar uma técnica não muito utilizada para este fim, que é a análise de Similaridade por Cosseno. Em toda a literatura que tivemos acesso, este cálculo de similaridade apenas foi utilizado, para auxiliar no cálculo de polaridades de termos, por (Esuli & Sebastiani, 2006), onde os significados das palavras do WordNet foram usados para calcular a similaridade semântica entre os termos avaliados, sabendo-se que termos similares tendem a ter a mesma orientação.

Sendo assim, neste trabalho, os seguintes métodos foram estudados:

1. Cálculo de polaridades através da análise de relações, proposto por (Godbole et al, 2007);
2. Cálculo de polaridades através do PageRank, algoritmo proposto para fins de análise de Internet por (Page et al, 1999) e, até onde pudemos observar, utilizado para este mesmo propósito apenas em (Esuli & Sebastiani, 2007);
3. Cálculo de polaridades através do SO-PMI, proposto por (Turney & Littman, 2003);
4. Cálculo de polaridades através de Similaridade por Cosseno, utilizado para este mesmo propósito apenas em (Esuli & Sebastiani, 2006);
5. Cálculo de polaridades através do *Google Similarity Distance*, proposto por (Cilibrasi & Vitányi, 2007);
6. Cálculo de polaridades através dos Coeficientes de Jaccard e Dice.

Como veremos mais adiante, os dois primeiros algoritmos analisam a similaridade semântica dos termos, ou seja, a semelhança que existe entre os significados das palavras analisadas, para calcular a polaridade de termos ainda desconhecidos. Já os demais, fazem uso da similaridade sintagmática, ou seja, avaliam a ocorrência de um determinado termo em relação aos demais já conhecidos para avaliar se são semelhantes e se podem, portanto, receber a mesma polaridade. Esses dois tipos principais de avaliar a similaridade entre termos são, inclusive, o que diferem o uso da Similaridade por Cosseno neste trabalho do uso dado em (Esuli & Sebastiani, 2006). Neste trabalho, esse algoritmo foi

utilizado para avaliar a similaridade sintagmática e não a similaridade semântica, como feito em (Esuli & Sebastiani, 2006). Essa diferenciação confere a este trabalho mais conhecimento obtido a partir do *corpus*, alinhando-se com o objetivo inicial do dicionário aqui gerado.

Vale ressaltar que não foi encontrada nenhuma ocorrência de uso desses métodos quando se trata do idioma português do Brasil. Além disso, também não foi possível encontrar nenhum trabalho que realizasse algum tipo de comparação entre o desempenho desses diferentes métodos ao classificar a orientação de termos.

4.4.1 Cálculo de polaridades através da análise de relações

Esta estratégia parte do princípio de que o dicionário previamente criado pode ser visto como um grafo, onde os nós são os termos e as arestas são as relações de sinonímia e antonímia existentes entre eles. Utilizada em (Godbole et al., 2007), ela foi escolhida para este trabalho por se basear em importantes pesquisas anteriores (Hatzivassiloglou & McKeown, 1997; Wiebe, 2000; Kim & Hovy, 2004) e por propor melhorias capazes de considerar a diferença de significados entre sinônimos de uma mesma palavra.

O cálculo de polaridades através da análise de relações se baseia nos caminhos traçados entre as palavras ao longo de um grafo. Conforme pudemos observar no capítulo anterior, diversas iniciativas de construção de léxicos fazem uso desse princípio ao partir de um pequeno conjunto de palavras *seed* e expandir esse grupo de termos a partir de relações obtidas através do WordNet. Em (Godbole et al, 2007), esse primeiro passo não foi diferente. No entanto, a estratégia de (Godbole et al, 2007) se diferencia das demais por considerar que a medida de similaridade entre dois termos é inversamente proporcional à distância entre eles no grafo, ou seja, quanto maior a distância, menos forte é a relação de sinonímia ou antonímia que os une. Podemos ver um exemplo desta afirmação na Figura 8, onde uma parte do grafo, gerado a partir do dicionário inicial criado neste trabalho, é apresentada. Nela podemos observar que o verbo “amar” pode ser considerado sinônimo de “maravilhar”. No entanto, o significado dos dois verbos, apesar de relacionados, já são distintos e não poderiam ser substituídos em uma frase sem prejudicar o sentimento expresso no texto original.

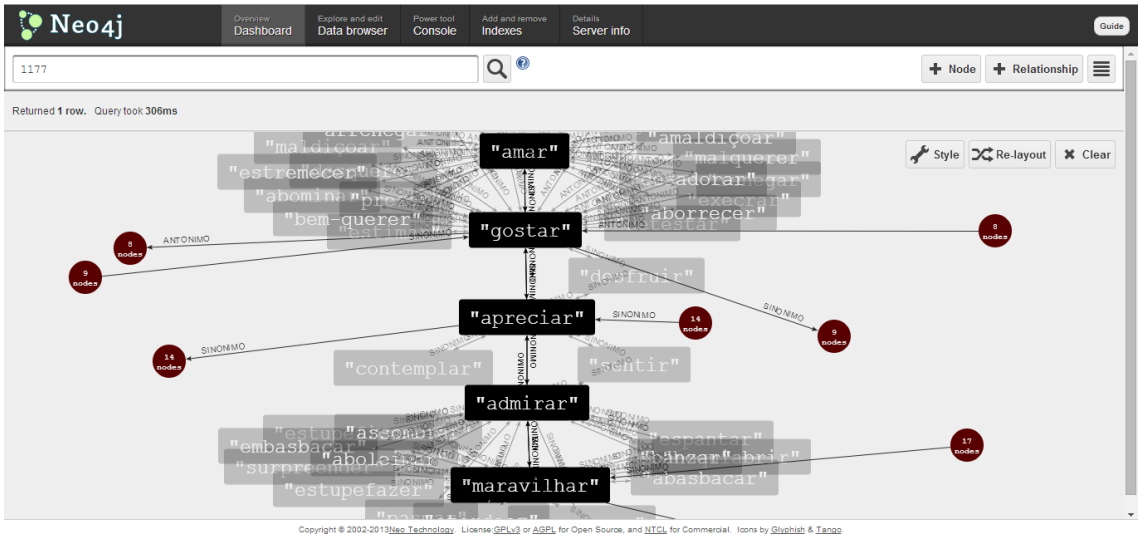

Figura 8 – Exemplo de alteração de significado dos sinônimos a medida que a distância aumenta entre eles no grafo.

Sendo assim, o algoritmo proposto em (Godbole et al, 2007), e utilizado neste trabalho, visa considerar esse decaimento da similaridade no momento de calcular a orientação repassada de uma palavra à outra. Para isso, ao analisarmos um caminho, consideramos as seguintes situações, assumindo que pol(A) denota a polaridade atribuída ao termo A:

- Se um nó A é filho direto de um nó B, pol(A) = pol(B), quando sinônimos, e pol(A) = (pol(B) * -1), quando antônimos;
- Se um nó A é filho indireto de um nó B, pol(A) = pol(B) * ($^{1}/_{C}{}^{d}$), onde C > 1 e d representa a profundidade de A em relação à B, no grafo. Para as relações de antonímia, a orientação final deve ser invertida;
- A polaridade final do nó A será a soma de todas as polaridades atribuídas a ele, ao longo dos diversos caminhos analisados.

Ao selecionar os caminhos existentes no grafo, não estamos livres de considerar caminhos considerados não confiáveis por possuírem um grande número de trocas de orientação ao longo do mesmo, conforme no exemplo da Figura 9 abaixo. Por isso, em um segundo momento do algoritmo, as polaridades provenientes de tais caminhos devem ser desconsideradas da soma da polaridade final do nó, influenciando neste cálculo apenas as polaridades originadas de caminhos considerados confiáveis. Neste trabalho, foram considerados caminhos não confiáveis aqueles que mudaram de orientação mais de três vezes, dado que a profundidade utilizada para selecionar tais caminhos variou de h=1 até h = 5, assim como no experimento original de (Godbole et al, 2007).

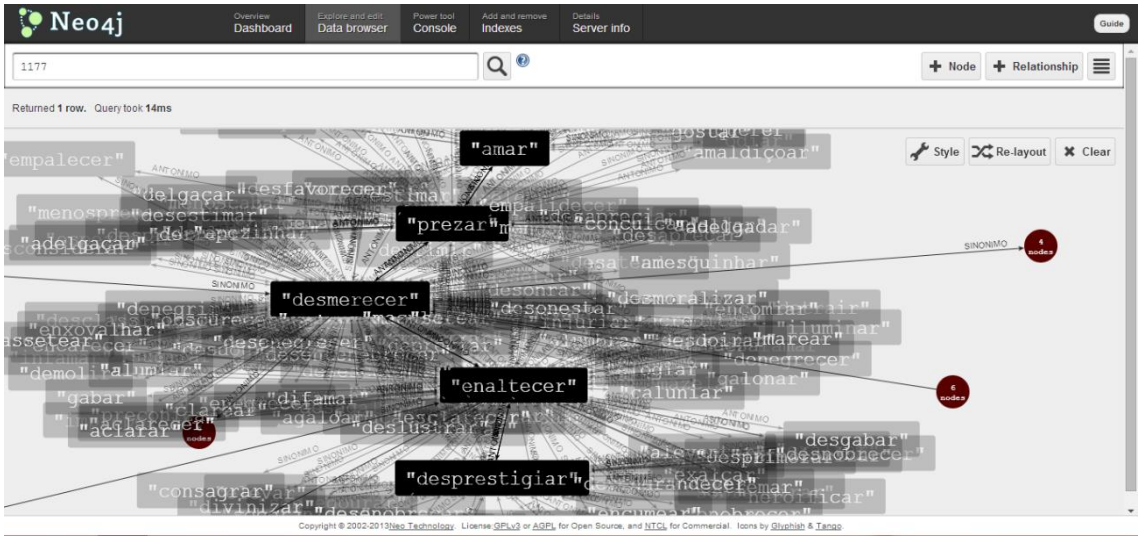

Figura 9 – Exemplo de caminho considerado não confiável e, portanto, descartado do cálculo final de polaridade do termo.

Esta estratégia para calcular a orientação dos termos se aplica apenas àqueles com polaridade ainda não calculada e que estejam presentes no dicionário inicial. Como neste trabalho não tratamos a questão de buscar sinônimos e antônimos dos termos extraídos do *corpus*, esta estratégia não possui informações suficientes sobre tais palavras para calcular suas polaridades. A opção de não tratar essas novas relações de sinonímia se deu por acreditarmos ser tema de um outro trabalho, apenas relacionado a este.

4.4.2 Cálculo de polaridades através do PageRank

O segundo algoritmo utilizado para calcular a polaridade dos termos também se baseia nas relações explicitadas através da visualização do dicionário inicial como um grafo e, portanto, também não abrange as palavras extraídas do *corpus*. Esta estratégia faz uso do PageRank, que, até onde pudemos verificar, só foi utilizado para este fim em (Esuli & Sebastiani, 2007), onde os autores se basearam num grafo criado a partir do WordNet para avaliar o nível de positividade e negatividade de cada palavra. Para isso, eles utilizaram o PageRank em dois momentos, primeiramente apenas com termos origem positivos e, depois, apenas com termos origem negativos. Tal experimento foi realizado com diferentes conjuntos de termos origem e obteve acurácia máxima de 67,5% para a classe positiva e 71,6% para a classe negativa.

Para entendermos um pouco melhor por que o PageRank pode ser aplicado para esta tarefa, precisamos, primeiramente, compreender seu comportamento e sua motivação. Este método foi criado por (Page et al, 1999) com o objetivo de classificar páginas da Internet de acordo com a importância daquele *site* para as pessoas. Para isso,

os autores buscaram analisar o grafo formado pela estrutura de *links* da *web*, onde cada nó do grafo representa uma página e cada aresta representa um *link* de uma página A para uma página B, e consideraram que as páginas mais importantes são aquelas que possuem maior número de apontamentos a partir de outras páginas também consideradas importantes.

Sendo assim, define-se formalmente o PageRank da seguinte maneira:

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u)$$

onde *R'(u)* representa o PageRank de um nó *u*, sendo N_v o número de *links* que partem do nó *v* e *R'(v)* o PageRank do nó *v*, onde *v* pertence ao conjunto de nós que apontam para *u*. A constante *c* aparece como um fator normalizador e *E(u)* indica a probabilidade de, aleatoriamente, um nó ser visitado a partir de um outro nó anterior aleatório. A existência de *E(u)* se faz necessária para evitar cálculos errôneos em casos em que o nó visitado não possui nenhuma aresta saindo dele. Tal situação causaria um acúmulo da classificação atribuída, já que a mesma não seria distribuída a nenhum outro nó. Portanto, *E(u)* evita esse problema e ainda funciona como um parâmetro de configuração do PageRank. Neste trabalho, o valor utilizado foi E(u) = 0.15, o mesmo proposto por (Page et al, 1999).

O funcionamento do PageRank nos permite imaginar que, se aplicado à descoberta de termos positivos e negativos, aquelas palavras que possuírem mais apontamentos oriundos de palavras positivas, podem também serem consideradas de orientação positiva. O mesmo se aplica para as demais classes de orientação. Tal suposição foi confirmada por (Esuli & Sebastiani, 2007) e também neste trabalho, que, apesar de se diferenciar levemente da estratégia de (Esuli & Sebastiani, 2007), também obteve bons resultados, conforme veremos no capitulo 5.

Outro ponto importante do PageRank é que, no algoritmo original, a probabilidade repassada por qualquer aresta saindo de um determinado nó é a mesma. Neste trabalho, essa probabilidade é alterada ao especificarmos os pesos das arestas. Esses pesos representam a orientação da palavra do nó origem multiplicada pelo valor da relação ali presente, sendo 1 quando é sinônimo e -1 quando é antônimo, quando o termo considerado possui polaridade advinda do SentiLex. Quando essa informação não está

presente, o peso da aresta não é informado e o algoritmo segue a probabilidade padrão estabelecida.

4.4.3 Cálculo de polaridades através do SO-PMI

Semantic Orientation from PMI (SO-PMI) (Turney & Littman, 2003) tem por objetivo identificar a orientação de uma determinada palavra. Para isso, ele utiliza o *Pointwise Mutual Information* (PMI), mais especificamente o PMI-IR (Turney, 2002), para avaliar o nível de associação semântica entre os termos e, a partir deste resultado, designar a polaridade calculada. O IR adicionado à sigla do PMI refere-se à Recuperação da Informação, representada pelas consultas a mecanismos de busca que fornecem os dados necessários para avaliar a associação existente entre os termos analisados. Como esse algoritmo se baseia apenas na análise dos documentos do *corpus*, ele nos permite calcular também as polaridades dos termos coletados dos textos através das regras de extração de adjetivos.

Para determinar a polaridade, o termo que se deseja investigar deve ser avaliado em relação à um conjunto de palavras que representam a classe positiva e um outro grupo representante da classe negativa. Cada uma das comparações, realizadas através do PMI, visa verificar se o termo analisado está associado ou não a uma das classes de orientação. Se a comparação for positiva, considera-se que as palavras avaliadas tendem a estar associadas. Se for negativa, assume-se a situação contrária. A polaridade será definida a partir da subtração dessas duas comparações, definindo o termo como positivo, quando SO-PMI(termo) > 0, e negativo, se SO-PMI < 0. Podemos definir formalmente o SO-PMI da seguinte forma:

$$Seja\ P_{termos} = conjunto\ de\ termos\ com\ orientação\ positiva$$

$$Seja\ N_{termos} = conjunto\ de\ termos\ com\ orientação\ negativa$$

$$SO - PMI(termo) = \sum_{ptermo \in P_{termos}} PMI(termo, ptermo) - \sum_{ntermo \in N_{termos}} PMI(termo, ntermo)$$

O PMI foi utilizado como função de associação por ser capaz de calcular "a força da associação semântica entre palavras" (Turney & Littman, 2003). Seu cálculo se baseia na probabilidade de coocorrência dos termos considerados e, tanto em (Turney, 2002) quando em (Turney & Littman, 2003), esses dados estatísticos foram obtidos através de

consultas à um mecanismo de busca na Internet. Dado que a intenção é verificar se as palavras analisadas têm maior probabilidade de ocorrerem juntas ou separadas, o número de resultados obtidos ao consultar um determinado termo no AltaVista é suficiente para fornecer esta informação. Sendo assim, podemos definir o PMI-IR da seguinte forma:

$$PMI(termo_1, termo_2) = \log_2 \left(\frac{\frac{1}{N} resultados(termo_1\ NEAR\ termo_2)}{\frac{1}{N}(resultados(termo_1)) \cdot \frac{1}{N}(resultados(termo_2))} \right)$$

onde *N* é o número total de documentos indexados e *resultados* representa a quantidade de documentos retornados para a consulta realizada.

A escolha do mecanismo de busca se deu devido à necessidade de uso do operador *NEAR*. Na época do estudo de Turney & Littman, o AltaVista era o único a permitir a utilização deste operador para que os termos consultados estivessem a uma distância de até 10 termos entre eles. Atualmente, nenhum mecanismo de busca na *web* oferece esta funcionalidade e esse foi um dos desafios da implementação deste algoritmo, conforme veremos no próximo capítulo.

Vale ressaltar que em (Turney & Littman, 2003) o algoritmo aqui apresentado também utilizou como função de associação a *Latent Semantic Analysis*, dando origem ao SO-LSA. No entanto, nos experimentos realizados em (Turney & Littman, 2003), o SO-PMI se mostrou superior ao SO-LSA, principalmente por ser altamente escalável, além de alcançar resultados suficientemente satisfatórios ao obter acurácia entre 82% e 97%, dependendo do conjunto de teste considerado.

4.4.4 Cálculo de polaridades através de Similaridade por Cosseno

O quarto algoritmo utilizado também se baseia em verificar a similaridade entre os termos para, assim, conseguir designar a polaridade dos mesmos. A Similaridade por Cosseno já foi amplamente utilizada em atividades de Recuperação da Informação para avaliar a similaridade entre documentos. Ela se baseia na representação dos documentos através de vetores, onde cada componente do vetor corresponde a cada um dos termos presentes no texto (Manning et al, 2008). Neste tipo de representação, todos os documentos são tratados como *bag of words* e dão origem a um conjunto de vetores que, dispostos em um espaço vetorial, permitem analisar a distância angular entre cada um

deles, característica utilizada pela Similaridade por Cosseno para atingir seu objetivo (Manning et al, 2008).

Aqui neste trabalho, nosso objetivo é utilizar a Similaridade por Cosseno para calcular o nível de similaridade entre os termos e, assim, inferir a sua orientação. Inicialmente a utilização desta estratégia para esta atividade era considerada em inédita, mas ao longo deste trabalho encontramos a iniciativa de (Esuli e Sebastiani, 2006), onde os significados das palavras do WordNet foram vetorizados para que fosse possível comparar a similaridade entre eles através deste cálculo. Termos com significados considerados semelhantes foram considerados de mesma orientação. Neste trabalho, a estratégia procura calcular a Similaridade por Cosseno existente entre um conjunto de documentos que contém o termo alvo e os conjuntos de documentos relacionados à termos positivos e à termos negativos. Para encontrar esses documentos, os conjuntos de palavras representativas de cada classe foram os mesmos utilizados pelo SO-PMI, como veremos mais adiante.

Após a seleção dos conjuntos de documentos de cada orientação e da vetorização deles, calcula-se a similaridade entre o termo em questão e cada um desses conjuntos da seguinte maneira:

$$sim(t, d_1) = \frac{\vec{V}(t) \cdot \vec{V}(d_1)}{\left|\vec{V}(t)\right| \left|\vec{V}(d_1)\right|}, onde\ \vec{x} \cdot \vec{y} = \sum_{i=1}^{M} x_i\, y_i \quad e \quad |\vec{V}(d)| = \sqrt{\sum_{i=1}^{M} \vec{V}_i^2(d)}$$

Calculada as similaridades entre o termo e as classes, aquela que obtiver maior semelhança será designada a classe do termo analisado.

Vale ressaltar que o vetor que representa os conjuntos de documentos possui, na realidade, o tf-idf de cada termo existente no texto considerado.

4.4.5 Cálculo de polaridades através de *Google Similarity Distance*

Assim como nos dois últimos algoritmos, aqui também experimentamos utilizar um outro cálculo de similaridade como insumo para a decisão sobre o repasse da polaridade entre duas palavras. Neste caso, o algoritmo testado é o proposto em (Cilibrasi & Vitanyi, 2007), que também se baseia na coocorrência dos termos em documentos, assim com o SO-PMI. A diferença entre os dois está no cálculo realizado para obter a

similaridade e no fato de que, na *Google Similarity Distance*, o mecanismo de busca utilizado para obter os dados necessários foi o Google.

A teoria por trás da *Google Similarity Distance* se baseia no fato de que a quantidade de páginas da Internet indexadas pelo Google é tão grande e o grupo de pessoas que escrevem esses textos é tão vasto e diversificado, que a maneira como as palavras se apresentam nos documentos investigados representam o uso real e as verdadeiras relações semânticas existentes entre essas palavras (Cilibrasi & Vitanyi, 2007). Sendo assim, se essas palavras tendem a ocorrer juntas, significa que elas possuem alta probabilidade de representar um mesmo significado. É com base neste conceito que as frequências dos termos nos documentos são utilizados para calcular a *Google Similarity Distance* através do que os autores chamam de *Normalized Google Distance* (NGD), de acordo com a seguinte fórmula:

$$NGD = \frac{max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - min\{\log f(x), \log f(y)\}}$$

onde f(x) representa a frequência da palavra X, f(y) represente a frequência da palavra Y e f(x,y) representa a frequência de X e Y em conjunto.

Neste trabalho, a NGD foi utilizada para calcular a similaridade entre um dado termo e conjuntos de termos representativos das orientações positiva e negativa. Aquela orientação considerada mais similar ao termo em questão estabelece a polaridade da palavra investigada.

4.4.6 Cálculo de polaridades através dos Coeficientes de Jaccard e de Dice

Buscando explorar um pouco mais os resultados que poderíamos obter ao aplicar a similaridade sintagmática à tarefa de classificar a orientação de termos, resolvemos testar o desempenho de medidas clássicas de similaridade quando aplicadas à mesma tarefa à qual submetemos os últimos algoritmos independentes das relações entre os termos. No caso, escolhemos testar os coeficientes de Jaccard e de Dice, de acordo com as fórmulas apresentadas abaixo:

$$Coeficiente\ de\ Jaccard = \frac{\sum_{i=1}^{t} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i y_i}$$

$$Coeficiente\ de\ Dice = \frac{2\ \sum_{i=1}^{t} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2}$$

onde x representa a frequência do termo x, y a frequência do termo y e xy a frequência deles em conjunto.

4.5 Conclusão

Exploramos neste capítulo todo o processo utilizado neste trabalho para a criação do dicionário em português do Brasil. Cada uma das estratégias utilizadas para calcular as polaridades do termos foi definida formalmente, além da fase inicial de criação do dicionário *seed*.

É importante dizer que todos os algoritmos aqui explicitados foram escolhidos por obterem bons resultados prévios, seja na execução desta mesma atividade ou de atividades similares, apresentando alto potencial para se ajustar ao objetivo deste trabalho. Como veremos no próximo capítulo, alguns desses bons resultados se confirmaram e, em alguns casos, puderam ser até melhorados, gerando ao final diferentes dicionários de polaridades, através de estratégias distintas.

5 Resultados

Conforme dito anteriormente, os algoritmos apresentados no capítulo anterior permitiram a criação de diferentes dicionários de polaridade. Esses dicionários não se diferenciam entre si pelos termos que os compõem, mas sim pelos valores das orientações calculadas, já que cada estratégia utilizada classificou cada palavra de maneira diferente. Por isso, o objetivo deste capítulo é apresentar e avaliar os resultados obtidos e, assim, investigar se os cálculos testados podem, realmente, ser aplicados na designação da polaridade de termos do português do Brasil. O foco principal é avaliar o desempenho de cada estratégia, principalmente em relação à sua acurácia.

Além disso, no capítulo anterior, pudemos verificar que, salvas as adaptações realizadas para este trabalho, os algoritmos selecionados já haviam sido utilizados em ao menos um outro trabalho. Sendo assim, neste capítulo, também faremos uma comparação dos resultados aqui obtidos com os reportados por outros pesquisadores, afim de analisar as vantagens, desvantagens e dificuldades de trabalhar com as bases de dados, ferramentas e dicionários *seeds* selecionados, e como essas escolhas influenciaram os resultados alcançados.

Por fim, avaliaremos também as classificações obtidas quando as diferentes estratégias são consideradas de forma conjunta. Para isto, na seção 5.1 deste capítulo discutiremos os dados e ferramentas utilizados na implementação dos dicionários; na seção 5.2, exploraremos os resultados obtidos através de cada estratégia; na seção 5.3, por fim, analisaremos tais resultados.

5.1 Ferramentas e bases de dados utilizadas

Antes de prosseguirmos aos resultados, é importante considerarmos os dados e as ferramentas utilizadas em cada uma das estratégias de cálculo de polaridades já explicitadas no capítulo anterior. Todas as implementações deste trabalho foram geradas com base na linguagem Java com o apoio de diferentes bibliotecas e ferramentas, que exploraremos a seguir. Além disso, durante todo o processo, também foi utilizado o banco de dados relacional MySQL para armazenamento dos termos do dicionário, das polaridades calculadas e demais informações de apoio aos algoritmos.

5.1.1 Ferramentas e bibliotecas

Uma das primeiras ferramentas incorporadas ao processo foi o GATE – *General Architecture for Text Engineering*, juntamente com os diversos recursos por ele oferecidos. O GATE é um projeto *open-source* de processamento de linguagem, que pode ser utilizado através de um ambiente próprio ou incorporado ao código Java, para que seja possível fazer uso das inúmeras funcionalidades de análise de textos nele existentes (GATE, 2014). Neste trabalho, ele foi utilizado na fase de extração de novos termos do *corpus* e os principais recursos usados foram o POS-Tagger e o JAPE.

Como vimos anteriormente, as regras de extração de termos eram baseadas na função gramatical exercida por cada palavra em uma determinada sentença. Por isso, a utilização de um POS-Tagger foi essencial para que, antes da aplicação das regras, fosse possível identificar quais termos estavam exercendo a função de adjetivo, principalmente. O GATE disponibiliza alguns POS-Taggers, porém todos eles são voltados para outros idiomas. Sendo assim, o GATE foi utilizado para, através de um plugin para o LingPipe – outra ferramenta de processamento de texto que faz uso de linguística computacional – , apoiar a execução de um POS-Tagger criado com foco no português do Brasil, implementado com *Hidden Markov Model* conforme descrito em (Manning & Schütze, 1999).

Além disso, o GATE também nos apoiou na aplicação das regras e descoberta de novos termos através do JAPE – *Java Annotation Patterns Engine*. O JAPE faz uso de expressão regular para encontrar determinados padrões no texto e achar a informação desejada (GATE, 2014). Sendo assim, as regras de extração foram descritas através do JAPE e, com base nas anotações geradas pelo POS-Tagger, os novos termos foram extraídos e incorporados ao dicionário.

Outra ferramenta importante foi o Neo4J, um banco de dados onde as informações são armazenadas e representadas através de grafos (Neo4J, 2014). Ele foi necessário para apoiar o algoritmo apresentado na subseção 4.4.1, que precisava visualizar os termos e seus relacionamentos através de um grafo para que fosse possível analisar as distâncias e os caminhos traçados entre as palavras. Sendo assim, todo o dicionário *seed* gerado foi incluído no Neo4J, onde os nós do grafo representaram os termos e as arestas corresponderam aos relacionamentos extraídos do TeP 2.0.

Além dessas ferramentas, algumas bibliotecas também foram utilizadas para apoiar a execução de parte dos algoritmos de cálculo de polaridades. A primeira delas foi o JUNG – *Java Universal Network/Graph Framework*, uma biblioteca Java que permite modelar, analisar e visualizar dados expostos como grafos ou redes (JUNG, 2010). Neste trabalho, o JUNG foi utilizado para apoiar a execução do algoritmo Page Rank, conforme apresentado na subseção 4.4.2, já implementado por esta biblioteca.

A segunda foi o Apache Lucene, biblioteca amplamente utilizada em algoritmos que envolvem a busca e recuperação de textos (LUCENE, 2012). No caso deste trabalho, o Lucene foi utilizado para apoiar a criação de um índice com base nos documentos do *corpus* analisado e, posteriormente, na execução de consultas em cima deste índice. Este recurso foi utilizado pelos algoritmos apresentados nas demais subseções da seção 4.4, já que optou-se por priorizar a análise do comportamento das palavras dentro do *corpus* tratado, não explorando a utilização de mecanismos de busca genéricos disponibilizados na Internet.

5.1.2 Bases de dados

Conforme citado anteriormente, a base de dados utilizada neste trabalho teve como origem o Twitter. Durante os meses de janeiro a março de 2012, mais de 550.000 *tweets* sobre o programa de televisão Big Brother Brasil foram coletados. A escolha desse tema se deu devido à popularidade deste programa, que possibilita alta geração de conteúdo em um curto espaço de tempo. Além disso, pudemos observar que grande parte desse conteúdo é de texto subjetivo, já que o assunto tende a gerar reações da população ligadas à simpatia que sentem pelos participantes deste *reality show*, o que acaba resultando em *tweets* que emitem opiniões e emoções sobre os acontecimentos e as pessoas do programa.

No entanto, o uso desta base apresentou algumas dificuldades no que diz respeito à análise do comportamento das palavras nos textos. Devido às suas características, já discutidas anteriormente, resolvemos avaliar os algoritmos dependentes do *corpus* também em outras bases para verificar se a estrutura dos *tweets* poderiam influenciar negativamente a performance dos mesmos. Sendo assim, foram testadas outras duas bases: ReLi, um *corpus* de cerca de 1600 resenhas de livros escritas em português do Brasil e manualmente rotuladas quanto à opinião nelas presente (Freitas et al, 2012), e

CETENFolha, uma base gerada a partir de textos do jornal Folha de São Paulo com cerca de 330.000 documentos e 24 milhões de termos (LINGUATECA, 2002).

Vale ressaltar que apesar do CETENFolha ser uma base de textos objetivos, isso não invalida a análise dos algoritmos através dos seus dados. Como os dois algoritmos dependentes do *corpus*, apresentados na subseções 4.4.3 e 4.4.4, tem como insumo a coocorrência e a frequência dos termos nos documentos, ainda é possível obter as informações necessárias para calcular as similaridades entre os termos através de textos que não emitem opinião. Esta característica do CETENFolha pode sim fornecer informações distintas das que seriam obtidas a partir de textos subjetivos, no entanto acreditamos que esta questão não afeta significativamente os resultados obtidos, como veremos mais à frente.

Por fim, é importante dizer que todos os textos analisados passaram por um pré-processamento simples de retirada de *stop words* e separação de conjunções como "do", "essas", "nessas", para que os documentos se adequassem melhor ao processo de POS Tagging.

5.1 Resultados

Para avaliar os resultados obtidos através de cada estratégia, primeiro precisamos compreender a maneira como esta análise foi realizada. Cada um dos algoritmos testados foi avaliado sob o método de avaliação cruzada 10-*fold*, onde 10 subconjuntos de termos são alternados entre grupos de treino e de teste. O conjunto de termos que deu origem a esses subconjuntos foi selecionado a partir do dicionário inicial gerado, de acordo com dois critérios: o termo precisava ser oriundo do TeP 2.0 para garantirmos que ele seria um termo do português do Brasil; e a orientação do termo extraída do Sentilex precisava ter sido manualmente etiquetada, para tentarmos reduzir a limitação imposta pelos resultados automáticos obtidos pelo algoritmo de classificação utilizado por (Silva et al, 2010). Tais filtros permitiram selecionar um conjunto de termos cujas características são apresentadas na Tabela 5 abaixo.

Tabela 5 – Quantidade e percentual de cada orientação dentre os termos utilizados na validação 10-fold

Positivos	Negativos	Neutros	Total
954	2252	429	3636
26,24%	**61,94%**	**11,82%**	**100%**

Vale ressaltar que nenhum termo extraído do *corpus* participa desta validação. Isso se deve principalmente por não nos permitirem avaliar todos os algoritmos utilizados, já que não temos informação sobre as relações de sinonímia e antonímia que eles possuem em relação aos demais termos do dicionário. De qualquer forma, na seção seguinte, teremos oportunidade de avaliar alguns dos termos extraídos e a polaridade à eles atribuída automaticamente.

5.1.1 Resultados do cálculo de polaridades através da análise de relações

Neste primeiro algoritmo, as relações entre os termos possuem extrema importância, pois é através delas que acontecerá o repasse de polaridades entre as palavras, de acordo com a distância entre elas. Aqui existem dois parâmetros que precisam ser considerados: a constante utilizada para calcular o decaimento da polaridade ao longo do caminho, que chamamos de C, e a constante que indica quantas vezes um caminho pode mudar sua orientação sem prejudicar a qualidade da informação repassada, que vamos chamar de O.

A variação de O está relacionada a quão rígidos seremos em relação à mudança de orientação entre as relações. Quando utilizamos O = 0, descartamos qualquer caminho que apresente mais de uma orientação em toda a sua extensão. Já quando consideramos O = 5, incluímos nos cálculos todo e qualquer caminho, independente das variações de orientação, já que estamos trabalhando com h = 5. Neste trabalho, utilizamos O = 3, permitindo que cada caminho tenha até três relações do tipo antônimo, responsáveis pela troca de orientação. Esse valor foi escolhido após verificarmos que o mesmo obteve o melhor resultado se comparado aos demais valores analisados, conforme podemos visualizar na Tabela 6. Para realizar esta análise, avaliamos a acurácia média do algoritmo de acordo com a variação de O, mantendo C = 2, já que quando C = 1 o decaimento não é considerado.

Tabela 6 – Acurácia média do algoritmo de análise de relações de acordo com a variação de O, mantendo C = 2.

O	Acurácia
0	66,0342%
1	69,6099%
2	73,9552%
3	**75,1659%**
4	74,9185%
5	74,9460%

Após fixarmos o valor de O, analisamos o valor de C da mesma maneira. Ao alterarmos o valor de C, modificamos quanto da polaridade de um termo será repassado a outro, ou seja, quanto é a contribuição de um termo sinônimo ou antônimo na polaridade de um outro termo. Assim como realizado com O, variamos os possíveis valores de C, para avaliar, de acordo com a acurácia do algoritmo, qual teria melhor desempenho. Esta análise, apresentada na Tabela 7, mostrou que C = 15 atinge os melhores resultados, mostrando que apenas $^{1}/_{15}$ da polaridade de um termo é suficiente para compor a polaridade de outro.

Tabela 7 – Acurácia média do algoritmo de análise de relações de acordo com a variação de C, mantendo O = 3.

C	Acurácia
2	75,1659%
3	76,1283%
4	77,0085%
5	77,1186%
10	77,6963%
15	**78,0265%**
20	78,0264%

Dadas estas observações, os valores de C e O foram ajustados para 15 e 3, respectivamente, para que pudéssemos realizar outras análises, baseadas nos resultados apresentados na Tabela 8 abaixo.

Tabela 8 – Precisão e revocação médios do algoritmo de análise de relações para cada orientação.

Acurácia: 78,0265%			
	Positivo	**Negativo**	**Neutro**
Precisão	69,4390%	**88,3311%**	28,5556%
Revocação	**87,0628%**	85,8005%	16,5972%

Primeiramente, podemos observar que este algoritmo se mostrou eficiente na classificação da orientação dos termos do dicionário, alcançando uma acurácia de 78%, como apresentado na Tabela 8, se igualando ao nível de corretude esperado de um humano, que acredita-se variar de 72% (Wiebe et al, 2006) a 85% (Golden, 2011). Além disso, o desempenho da classe negativa é superior, o que pode ser explicado pela parcela significativamente maior de termos negativos dentre as palavras consideradas. No entanto, podemos observar que o desempenho da classe positiva também é satisfatório, principalmente quando avaliada em relação à revocação desta orientação. Já a classe neutra apresenta resultados realmente aquém do desejado, possível consequência do baixo percentual de termos neutros no dicionário.

Por fim, se compararmos o desempenho obtido com o reportado por (Godbole et al, 2007), veremos que os resultados aqui descritos estão bastante alinhados com os apresentados pelo estudo original, o que demonstra que o algoritmo original não perdeu desempenho ao trabalhar com as bases utilizadas neste trabalho. Em (Godbole et al, 2007), o algoritmo foi avaliado de acordo com as medidas de precisão e revocação, sendo que eles também observaram que o melhor desempenho é alcançado quando O = 3. Na Tabela 9 abaixo são apresentados os resultados obtidos pelos autores. No estudo original, os resultados para a classe neutra não foram reportados.

Tabela 9 – Resultados reportados por (Godbole et al, 2007).

	Positivo	Negativo
Qtd. termos	344	386
Precisão	84%	82,7%
Revocação	64,2%	69,4%

5.1.2 Resultados do cálculo de polaridades através do Page Rank

O Page Rank é o segundo dos dois algoritmos que se baseiam nas relações de antônimos e sinônimos expostas através de um grafo para calcular as polaridades dos termos. Nesse algoritmo, todos os repasses são baseados exclusivamente nas relações e na probabilidade de um determinado nó chegar até outro, de acordo com o que foi descrito na seção 4.4.2 do capítulo anterior.

O Page Rank atingiu uma acurácia média satisfatória de 73,8027%. Uma desvantagem do Page Rank em relação ao algoritmo anterior é que ele não se mostrou adequado para a classificação de termos neutros, obtendo médias de 0% para esta classe, como podemos observar na Tabela 10. Isso se deve à natureza do algoritmo que assume que cada nó tem probabilidade > 0 de ser visitado, sendo assim, nenhum nó ao final do algoritmo é associado à probabilidade = 0, o que, para este trabalho, indicaria o termo ser neutro.

Tabela 10 – Precisão e revocação médios do algoritmo Page Rank para cada orientação.

Acurácia: 73,8027%			
	Positivo	Negativo	Neutro
Precisão	54,0676%	**89,9627%**	0%
Revocação	**92,44%**	80,0128%	0%

O desempenho superior da precisão da classe negativa se justifica pelas mesmas razões explicitadas em relação ao algoritmo de análise de relações. Por existir um número maior de palavras negativas, obtém-se maior informação sobre essa classe, aumentando

as chances de acerto quando encontramos novos termos pertencentes à esta orientação. Neste caso, a classe positiva já foi mais prejudicada na sua precisão, atingindo 54,0676%, porém a revocação para esta orientação foi muito boa, mostrando que o algoritmo é capaz de detectar corretamente a grande maioria das palavras positivas.

Conforme dito anteriormente, só observamos o uso do Page Rank para auxiliar no cálculo de polaridades em (Esuli & Sebastiani, 2007), onde os resultados obtidos foram avaliados de acordo com a medida do coeficiente de Kendall. Ao interpretarmos esses resultados, verificamos que foi alcançada uma taxa de acerto de 67,5% para termos positivos e 71,6% para termos negativos. No caso de (Esuli & Sebastiani, 2007), as relações entre os termos foram obtidas através do WordNet e as polaridades *seed* através do SentiWordNet.

A partir destes resultados, podemos concluir que o uso das informações do dicionário *seed* adotado neste trabalho foram mais adequadas para o idioma tratado, conseguindo até alcançar desempenho superior em relação à outra iniciativa de uso do Page Rank para a classificação da orientação de termos.

5.1.3 Resultados do cálculo de polaridades através do SO-PMI

Até o momento analisamos os resultados obtidos para aqueles algoritmos dependentes das relações existentes entre os termos. Com o SO-PMI, assim como no caso da Similaridade por Cosseno que veremos mais adiante, podemos calcular a polaridade não apenas dos termos do dicionário *seed*, mas também das novas palavras extraídas do *corpus* analisado. No entanto, é importante dizer que, para efeito de comparação, os resultados aqui apresentados se baseiam no mesmo grupo de termos utilizados até agora para avaliação dos demais algoritmos, mantendo a consistência da base utilizada para a validação cruzada.

Conforme explicitado na seção 4.4.3, o SO-PMI se baseia na comparação de um termo que se deseja investigar e um conjunto de palavras que representam uma determinada orientação. No trabalho original realizado por (Turney & Littman, 2003), dois conjuntos foram utilizados para representar cada uma das classes tratadas. Esses conjuntos são compostos, cada um, por 7 palavras independentes de domínio, ou seja, termos que mantém sua orientação independente do significado de outros termos que os acompanham. São eles:

G_POS = {good, nice, excelent, positive, fortunate, correct, superior}

G_NEG = {bad, nasty, poor, negative, unfortunate, wrong, inferior}

Neste trabalho, ao avaliar este algoritmo, primeiramente optamos por manter os mesmos grupos de palavras. Sendo assim, através de uma tradução livre, que levou em consideração também a ocorrência dessas palavras no *corpus*, obtivemos os seguintes conjuntos representativos das classes positiva e negativa para o português do Brasil:

G_POS = {bom, legal, excelente, positivo, feliz, certo, melhor}

G_NEG = {ruim, chato, pobre, negativo, triste, errado, pior}

Os experimentos com esses conjuntos nos permitiram obter os resultados apresentados na Tabela 11. Vale ressaltar que, diferente do estudo original de (Turney & Littman, 2003), aqui as buscas foram realizadas através de um índice Lucene criado com base nos documentos do *corpus*, os *tweets*. Essa escolha foi realizada por querermos obter informações sobre as palavras da maneira mais contextualizada possível. Essa opção nos trouxe também algumas desvantagens, como o fato de diversas palavras analisadas não coocorrerem com os termos de G_POS e G_NEG nos documentos, impossibilitando esta análise em relação a 63,6139% dessas palavras. Isso ocorre, principalmente, devido ao fato de muitas palavras que compõem o conjunto de teste utilizado não estarem presentes em textos como *tweets*. Outro possível fator influenciador é o fato de os textos analisados serem bastante curtos, o que diminui a chance de duas palavras coocorrerem em um mesmo documento. O impacto direto destas características pode ser visto na acurácia média de 50,7232%, alcançada quando os grupos representativos das classes foram os apresentados acima.

Tabela 11 – Precisão e revocação médios do algoritmo SO-PMI para cada orientação, quando utilizado o *corpus* de *tweets*.

Acurácia: 50,7232%			
	Positivo	**Negativo**	**Neutro**
Precisão	38,2382%	**59,9323%**	0%
Revocação	49,9988%	**61,7211%**	0%

Em vista destes resultados e dos problemas apresentados, resolvemos avaliar o SO-PMI utilizando documentos de tamanho maior e com escrita menos informal, em busca de aumentar o nível de coocorrência entre as palavras avaliadas e os termos de G_POS e G_NEG e, assim, diminuir a quantidade de termos do conjunto de testes que

não receberam classificação. Para isso, o mesmo experimento foi realizado com a base ReLi e com o CETENFolha, conforme apresentado nas Tabelas 12 e 13, respectivamente.

Tabela 12 – Precisão e revocação médios do algoritmo SO-PMI para cada orientação, quando utilizado o *corpus* ReLi.

Acurácia: 49,7845%			
	Positivo	**Negativo**	**Neutro**
Precisão	30,7496%	**57,1600%**	0%
Revocação	28,5898%	**71,4989%**	0%

Tabela 13 – Precisão e revocação médios do algoritmo SO-PMI para cada orientação, quando utilizado o *corpus* CETENFolha.

Acurácia: 55,2012%			
	Positivo	**Negativo**	**Neutro**
Precisão	40,0040%	**61,2829%**	0%
Revocação	38,4753%	**76,1243%**	0%

Ao utilizamos a base CETENFolha, tivemos dois ganhos: redução do número de palavras não classificadas por falta de coocorrência nos textos, como pode ser observado na tabela 14; e melhora da acurácia em aproximadamente 5%, se comparado com os resultados obtidos na base original de *tweets*. Uma possível justificativa para o resultado inferior obtido pelo *corpus* ReLi está no tamanho da base que, apesar de apresentar textos mais extensos, possui um número de documentos muito inferior à quantidade de *tweets*, o que reduz bastante a quantidade de informação existente sobre as palavras analisadas, como podemos observar através do percentual de polaridades não calculadas apresentado na tabela abaixo.

Tabela 14 – Percentual de polaridades não calculadas com o SO-PMI, de acordo com a base utilizada.

Base	Percentual de polaridades não calculadas
Twitter	63,6139%
ReLi	83,3058%
CETENFolha	32,2057%

Apesar das melhorias alcançadas, resolvemos investigar se ainda seria possível melhorar os resultados obtidos. Sendo assim, optamos por alterar os conjuntos de palavras usadas como referência de cada classe para que fossem compostos pelos termos que mais ocorrem nos documentos das bases analisadas, com o objetivo de verificar se ao utilizar termos mais frequentes para compor o G_POS e o G_NEG, o número de coocorrências encontradas nos documentos também aumentaria. Para isso, calculamos a frequência de cada termo nos textos analisados e utilizamos novamente a base CETENFolha, já que a

mesma obteve melhor desempenho, para avaliar o algoritmo, verificando qual seria o ganho na acurácia se trabalhássemos com grupos de palavras de diferentes tamanhos, que chamaremos de N. Esta análise pode ser vista na Tabela 15 abaixo.

Tabela 15 – Acurácia média do algoritmo SO-PMI para a base CETENFolha, de acordo com a variação de N.

N	Acurácia
7	45,7922%
10	50,4623 %
15	**52,4566%**
20	51,3960 %

Através desta análise, pudemos observar que mesmo alterando os grupos representativos de cada classe e os seus tamanhos, não foi possível obter um resultado superior ao obtido com os grupos de referência utilizados pelo SO-PMI original. Acreditamos que isso se deve ao fato de alta frequência no *corpus* não implicar em alta coocorrência com os termos do conjunto de treino e teste e, por isso, o desempenho obtido anteriormente foi melhor. No entanto, observamos que ao utilizar e aumentar o tamanho desses grupos retirados do *corpus*, conseguimos diminuir a quantidade de termos com polaridade não calculada, conforme apresentado na Tabela 16 abaixo. Este dado demonstra que a utilização de G_POS e G_NEG composto por termos frequentes no *corpus*, permite aumentar a quantidade de termos etiquetados, mas não necessariamente nos traz mais informação sobre a coocorrência destes termos.

Tabela 16 – Percentual de polaridades não calculadas com o SO-PMI para a base CETENFolha, de acordo com a variação de N.

N	Percentual de polaridades não calculadas
7	**16,0616%**
10	**15,8141%**
15	**15,5391%**
20	**15,0440%**

Vale ressaltar que ao utilizarmos estes conjuntos de G_POS e G_NEG para a base de *tweets* e para o ReLi, os resultados foram melhores do que os obtidos anteriormente, porém não superiores ao obtido com o CETENFolha ao trabalharmos com os conjuntos originais de G_POS e G_NEG. Essa diferença de desempenho entre as bases se dá justamente pelo comportamento distinto dos termos em cada uma delas e pela diferente frequência das coocorrências, dado o conjunto de termos analisado neste experimento.

Se comparado com os resultados obtidos por (Turney & Littman, 2003), que chegaram a atingir 87% de acurácia quando utilizada toda a base de teste, o resultado aqui reportado ficou abaixo do que era esperado alcançar ao utilizar este algoritmo. No entanto, devemos considerar que aqui as informações sobre as coocorrências das palavras foram obtidas dentro do próprio *corpus*, tornando os resultados mais influenciados pelo domínio analisado, porém mais limitados às informações presentes nos documentos analisados. Sendo assim, os resultados obtidos, principalmente através do CETENFolha, demonstram que este algoritmo, quando executado da maneira aqui proposta, possui potencial para calcular a polaridade de termos em português do Brasil. Sendo assim, para obter resultados mais satisfatórios, é preciso apenas adequar a base utilizada aos termos avaliados de forma que seja possível extrair mais informações sobre a coocorrência das palavras nos documentos considerados.

5.1.4 Resultados do cálculo de polaridades através de Similaridade por Cosseno

Conforme dito anteriormente, a Similaridade por Cosseno também é um algoritmo que não depende das relações de sinonímia e antonímia entre as palavras. Na verdade, ele se baseia na distância entre os vetores que representam os termos avaliados para dizer se eles apresentam ou não alguma similaridade. Neste trabalho, essa informação é utilizada para avaliar se eles devem ser classificados como termos de uma mesma orientação.

O primeiro passo foi identificar quais termos serviriam como representantes de cada uma das classes. Para isso, utilizamos estratégia semelhante à utilizada no algoritmo SO-PMI, utilizando como termos referência das orientações os termos fixos designados por (Turney & Littman, 2003) e traduzidos livremente por nós. Essa estratégia foi utilizada por não termos a informação de quais documentos eram positivos e quais eram negativos em todas as bases. Por isso, precisamos optar por um modo de classificar esses documentos e escolhemos seguir por este caminho. Além disso, para este algoritmo, optamos por não trabalhar com outros conjuntos de termos referência por identificarmos que o conjunto utilizado conseguiu abranger um número considerável de palavras presentes no *corpus* através da recuperação dos documentos a elas relacionados, não se fazendo necessário alterar esses termos com o objetivo de melhorar o alcance dos vetores gerados para o algoritmo.

O segundo passo era representar esses conjuntos de termos como vetores. Por isso, para construir cada um dos vetores utilizados em cada comparação utilizamos o seguinte algoritmo:

1. Para cada termo representativo de cada classe, buscar os documentos do *corpus* a ele associados.
2. Construir um índice baseado apenas nesses documentos recuperados do *corpus* geral.
3. Para cada documento retornado, buscar as palavras existentes que não sejam *stop words*.
4. Para cada palavra encontrada, calcular o tf-idf dela no *corpus* gerado especificamente para a orientação em questão, através do passo 2.
5. Se o tf-idf dela for diferente de 0, colocá-la no vetor resultante.

Esse passo a passo nos permitiu criar os vetores representantes da classe positiva, da classe negativa e também do vetor que estava sendo investigado, isoladamente, para só então partir para a etapa seguinte.

O terceiro passo consistiu em unir esses 3 vetores iniciais de forma a gerar outros 3 vetores que contivessem informação sobre todas as palavras encontradas. A diferença entre cada um desses vetores está no valor do tf-idf de cada termo. Essas diferenças entre as frequências das palavras em cada vetor é suficiente para fazer uso da fórmula apresentada na seção 4.4.4 e conseguir calcular a distância entre cada um dos vetores. O resultado final indica que se a distância entre o vetor do termo investigado e o vetor da classe positiva for menor do que a distância entre ele e a classe negativa, ele pode ser considerado um termo positivo. Caso contrário, ele será considerado um termo negativo.

Ao aplicar este algoritmo à base de *tweets*, utilizando os termos referência citados anteriormente, obtivemos os valores médios reportados na Tabela 17. No entanto, percebemos que o número de polaridades não calculadas foi considerável, atingindo 33,5%. Esse percentual de termos não considerados deve-se ao fato de não existirem documentos no *corpus* geral que contivessem tais palavras e, portanto, não foi possível gerar vetores que as representassem. Sendo assim, para buscar investigar melhor este algoritmo, realizamos testes com as outras bases já citadas. Os teste com o ReLi se mostraram pouco produtivos, alcançando uma acurácia de apenas 39,7792% e com um

percentual de 51,2% de termos não classificados, também por não conseguirmos encontrar documentos que contivessem os termos presentes nos conjuntos de treino e teste utilizados. Já os resultados com a base CETENFolha foram mais promissores, conforme apresentado na Tabela 18.

Tabela 17 – Precisão e revocação médios do algoritmo Similaridade por Cosseno para cada orientação, quando utilizado o *corpus* de *tweets*.

Acurácia: 52,2068%			
	Positivo	**Negativo**	**Neutro**
Precisão	50,1905%	**58,4009%**	10,9973%
Revocação	3,6570%	**84,9110%**	11,9093%

Tabela 18 – Precisão e revocação médios do algoritmo Similaridade por Cosseno para cada orientação, quando utilizado o *corpus* CETENFolha.

Acurácia: 58,2991%			
	Positivo	**Negativo**	**Neutro**
Precisão	34,2857%	**60,1106%**	4,2063%
Revocação	0,6793%	**95,1385%**	4,2662%

Ao observarmos os resultados reportados acima podemos verificar que a Similaridade por Cosseno possui potencial para atuar também nesta atividade, necessitando apenas de um *corpus* que seja capaz de oferecer informações suficientes sobre os termos investigados. Ao utilizarmos a base do ReLi, a acurácia obtida foi bem abaixo do esperado e, ao observar a quantidade de termos não classificados, fica claro que o tamanho da base impactou negativamente no desempenho do algoritmo. Já com a base do CETENFolha, menos de 10% dos termos analisados não tiveram suas polaridades calculadas, o que demonstra que quanto maior for o *corpus* e quanto mais informação ele tiver, melhor será o desempenho da Similaridade por Cosseno.

5.1.5 Resultados do cálculo de polaridades através de *Google Similarity Distance*

Devido aos resultados obtidos nas estratégias anteriores, e para efeitos de comparação com os demais resultados, resolvemos adotar o CETENFolha como *corpus* padrão para os 3 últimos algoritmos. Essa escolha se deu por avaliarmos que esta base foi a que melhor se adequou ao objetivo de avaliar o potencial de cada algoritmo para realizar a tarefa em questão, dado que ele é o que fornece mais informações sobre as palavras consideradas nos experimentos, e também por ter obtido desempenho superior nos algoritmos já avaliados.

Assim como nos dois últimos algoritmos anteriores, a estratégia de calcular a similaridade entre um determinado termo e um conjunto de palavras previamente classificadas como positivas e negativas também se manteve com a *Google Similarity Distance*. Nesse caso, voltamos a explorar o uso de termos do *corpus*, já que voltamos a tratar da frequência isolada dos termos selecionados e, portanto, criar conjuntos de termos referência mais frequentes dentro do domínio analisado poderia influenciar o resultado final. Através de testes similares aos realizados com o SO-PMI, observamos que a melhor acurácia é alcançada ao utilizar grupos de 10 termos mais frequentes no *corpus*, conforme podemos observar nas Tabelas 19 e 20 abaixo.

Tabela 19 – Acurácia média do algoritmo NGD para a base CETENFolha, de acordo com a variação de N ou com o uso dos termos referência fixos.

N	Acurácia
Fixos	47,3422%
7	54,0848%
10	**55,8676%**
15	52,6564%
20	51,1904%

Tabela 20 – Precisão e revocação médios do algoritmo NGD para cada orientação, quando utilizado o *corpus* CETENFolha.

Acurácia: 55,8676%			
	Positivo	**Negativo**	**Neutro**
Precisão	32,9554%	**60,7649%**	0%
Revocação	20,3141%	**84,5633%**	0%

As similaridades foram calculadas para cada par de termos e, ao final, foi calculada a média desses valores e a polaridade correspondente ao grupo de termos mais similar foi atribuída ao termo investigado.

A acurácia obtida através do NGD pode ser considerado um bom resultado, ainda mais se considerarmos o baixo custo dos cálculos realizados. No entanto, a Similaridade por Cosseno ainda apresentou resultado superior. Outro ponto a se observar é que, assim como no SO-PMI, neste trabalho não foi utilizada o mesmo mecanismo de busca do experimento original realizado em (Cilibrasi & Vitanyi, 2007). Todas as frequências foram obtidas internas ao corpus do CETENFolha, o que aumenta o nível de contextualização, mas diminui a quantidade de informação sobre os termos, conforme já dito anteriormente.

É importante citar que, para este algoritmo, pudemos observar que o *corpus* oriundo do Twitter obteve resultado superior ao obtido através do *corpus* do CETENFolha. Esse e outros resultados podem ser visualizados no Apêndice I deste trabalho.

5.1.6 Resultados do cálculo de polaridades através dos Coeficientes de Jaccard e de Dice

Os Coeficientes de Jaccard e de Dice também foram testados em relação aos conjuntos de termos referência, assim como os demais algoritmos de similaridade sintagmática. Assim como com o *Google Similarity Distance*, estes dois também foram avaliados a partir do cálculo da média da similaridade entre uma determinada palavra e cada um dos termos referência de uma determinada classe. Por serem muito similares, já que analisam o mesmo aspecto sobre as palavras, diferenciando apenas o cálculo utilizado, resolvemos abordar os resultados encontrados em uma mesma seção.

Para o Coeficiente de Jaccard, o melhor resultado, com base no *corpus* CETENFolha, foi obtido quando os conjuntos representativos de cada orientação eram compostos por 15 termos. Tal resultado pode ser visualizado na Tabela 21 abaixo.

Tabela 21 – Precisão e revocação médios do algoritmo de Jaccard para cada orientação, quando utilizado o *corpus* CETENFolha.

Acurácia: 56,9737%			
	Positivo	**Negativo**	**Neutro**
Precisão	23,6772%	**59,4486%**	0%
Revocação	5,6068%	**92,5219%**	0%

Por serem muito parecidos, os resultados obtidos com o Coeficiente de Dice foram muito próximos aos de Jaccard, sendo alcançados com a mesma configuração dos grupos de termos referência do algoritmo anterior. Tal resultado pode ser visualizado na Tabela 22 abaixo.

Tabela 22 – Precisão e revocação médios do algoritmo de Jaccard para cada orientação, quando utilizado o *corpus* CETENFolha.

Acurácia: 56,7752%			
	Positivo	**Negativo**	**Neutro**
Precisão	23,3071%	**59,4653%**	0%
Revocação	5,9301%	**92,0069%**	0%

5.1.7 Combinação dos resultados

Como uma última estratégia de cálculo das polaridades resolvemos analisar o desempenho alcançado ao combinar os resultados obtidos previamente. Claramente alguns algoritmos tiveram melhores resultados que outros, principalmente se considerarmos os acertos de cada orientação de forma diferenciada. Sendo assim, para tentar buscar resultados ainda mais satisfatórios, foi realizada a combinação dessas classificações de forma a explorar o que cada algoritmo trouxe de melhor para o dicionário.

Para realizar esta análise foram utilizadas duas estratégias: combinar os resultados por meio de voto e soma. Essas estratégias foram escolhidas por representarem métodos adequados para alcançar, através da combinação dos resultados iniciais, melhores acurácias e, segundo (Xia et al, 2011), serem algoritmos que alcançam resultados satisfatórios com menor custo de processamento. Cada uma destas estratégias foi avaliada da mesma maneira que os classificadores individuais foram avaliados para que pudéssemos realizar uma comparação dos resultados obtidos por cada um deles, como veremos mais adiante.

5.1.7.1 Por Voto

Nesta estratégia o objetivo é verificar, dentre as 4 polaridades recebidas para cada termo, qual delas foi a mais votada pelos algoritmos utilizados para a classificação (Xia et al, 2011). Sendo assim, na Tabela 23 abaixo, podemos observar os resultados obtidos através desta técnica de análise dos resultados em conjunto.

Tabela 23 – Valores médios para a combinação por voto.

Acurácia: 66,7219%			
	Positivo	**Negativo**	**Neutro**
Precisão	63,2076%	**69,9472%**	22,9474%
Revocação	38,0934%	**90,0579%**	8,2413%

Os resultados acima apresentados demonstram que esta estratégia consegue alcançar resultados até superiores à alguns dos algoritmos previamente apresentados, elevando em até pouco mais de 12% a acurácia daqueles que obtiveram pior desempenho. Sendo assim, ela pode ser considerada um boa estratégia para melhorar os resultados obtidos individualmente, principalmente quando todo os conhecimento sobre os termos é extraído diretamente do *corpus* e o mesmo não fornece informações suficientes para os cálculos anteriores.

5.1.7.2 Por Soma

Nesta estratégia, somam-se as 4 classificações atribuídas aos termos para gerar a classificação final obtida através da combinação destes resultados. Segundo (Xia et al, 2011), esta técnica é semelhante à realizar a média dos resultados, porém com melhores resultados por ser capaz de lidar melhor com erros de estimativa. Na tabela 24 abaixo, podemos observar os resultados obtidos através deste método.

Tabela 24 – Valores médios para a combinação por soma.

Acurácia: 66,4194%			
	Positivo	**Negativo**	**Neutro**
Precisão	57,2345%	**69,9862%**	17,4924%
Revocação	38,7038%	**90,2316%**	3,4048%

Os resultados acima apresentados demonstram que esta estratégia alcança resultados bastante similares aos da Combinação por Voto, porém ficando ligeiramente abaixo da primeira. Mesmo assim, ela também poderia ser considerada um boa estratégia para melhorar os resultados obtidos individualmente.

5.2 Análise dos Resultados

Nesta seção, analisaremos os resultados reportados acima e realizaremos uma comparação dos mesmos, com base nas Figuras 10 e 11, apresentadas abaixo.

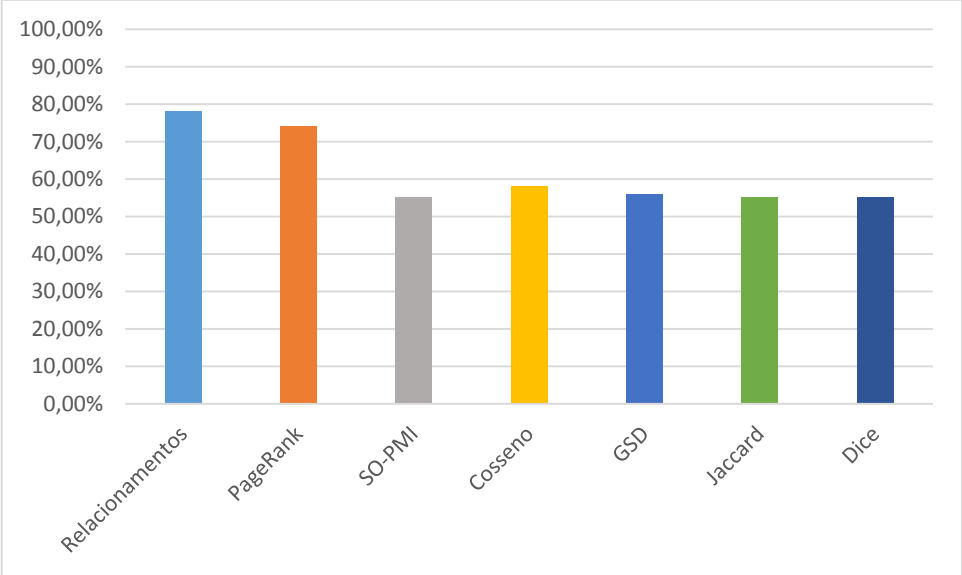

Figura 10 – Comparação entre as acurácias máximas obtidas através de cada algoritmo de cálculo de polaridades, tomando como base o *corpus* CETENFolha.

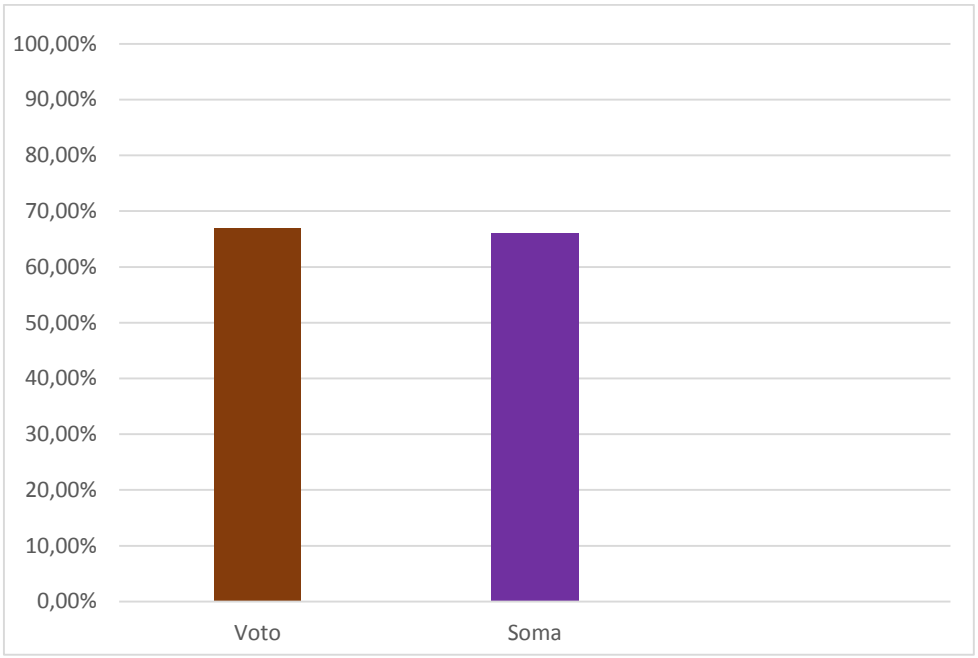

Figura 11 – Comparação entre as acurácias obtidas através de cada algoritmo de combinação das polaridades.

Analisando os resultados expostos acima, fica claro que os algoritmos independentes do *corpus*, que se baseiam no conhecimento das relações de sinonímia e antonímia dos termos, foram muito superiores aos demais. Esse resultado demonstra a importância dessa informação para o objetivo deste trabalho e sugere que, se soubermos as relações que existem entre os termos, conseguiremos gerar dicionários de polaridades com altíssima precisão e dependentes de domínio, já que as palavras extraídas dos textos também poderiam ser avaliadas desta maneira caso soubéssemos suas relações com os outros termos oriundos dos dicionários *seed*.

Os algoritmos dependentes da análise dos documentos do *corpus*, no entanto, apresentaram desempenhos piores. Ao limitar o SO-PMI aos documentos do *corpus*, e não utilizar a Internet como feito no experimento original, o seu desempenho caiu muito em relação ao reportado pelos seus autores. Em contrapartida, ficou claro para nós que o fato de muitos termos considerados nesta avaliação não aparecerem em conjunto com os termos do grupo de referência, prejudicou consideravelmente o desempenho deste algoritmo e, talvez, esta seja a razão para (Turney & Littman, 2003) terem buscado esse conhecimento na Internet e não em uma base específica. Já a Similaridade por Cosseno demonstrou também ser altamente dependente das informações presentes na base, porém, ainda assim, superou não só o SO-PMI como os demais algoritmos utilizados sob estas condições. Sendo assim, acreditamos que este algoritmo possui potencial para auxiliar na

atividade do cálculo de polaridades de um dicionário, precisando apenas contar com uma base que possua informação suficiente sobre os termos que serão avaliados.

Apesar de os algoritmos de combinação melhorarem os resultados obtidos pelos algoritmos dependentes do *corpus*, ainda assim eles tiveram um resultado inferior aos algoritmos baseados nas relações dos termos. Sem dúvida, as estratégias de similaridade semântica se destacaram, mostrando ser esse tipo de análise o caminho principal para a construção de um dicionário de polaridades, seja ele contextualizado ou não. Além do desempenho superior, essas estratégias também mostraram ser bastante consistentes, atingindo uma taxa de concordância nas classificações de 87,13%, como podemos ver na Figuras 12 e 13 abaixo.

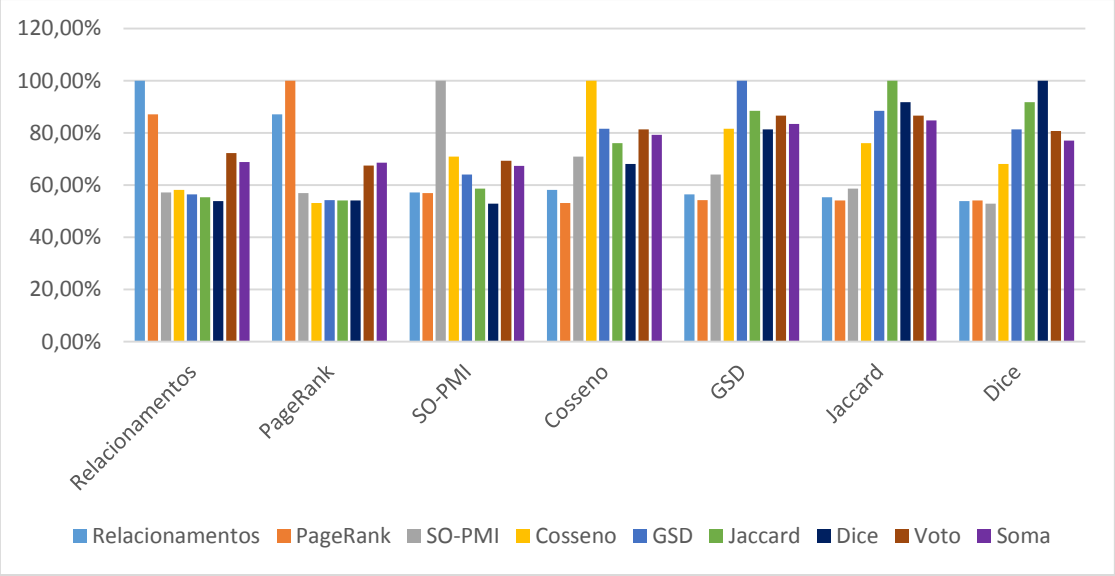

Figura 12 – Percentuais de concordância entre as classificações obtidas a partir de cada algoritmo.

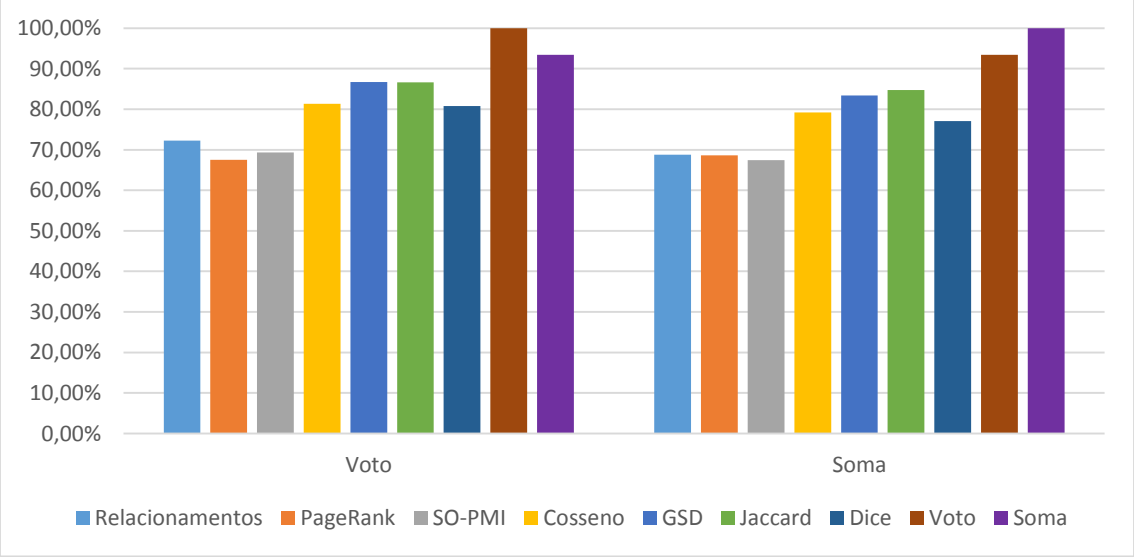

Figura 13 – Percentuais de concordância entre as classificações obtidas a partir de cada algoritmo.

Outro fator que vale ser analisado é a cobertura do cálculo das polaridades dos termos do TeP 2.0. Como dito anteriormente, no dicionário *seed*, apenas 11% do TeP 2.0 possui informação sobre sua orientação oriunda do SentiLex. Após realizar os cálculos propostos, o número de termos com polaridade calculada chegou a alcançar 100% para alguns dos algoritmos utilizados, como podemos observar na Figura 14 abaixo.

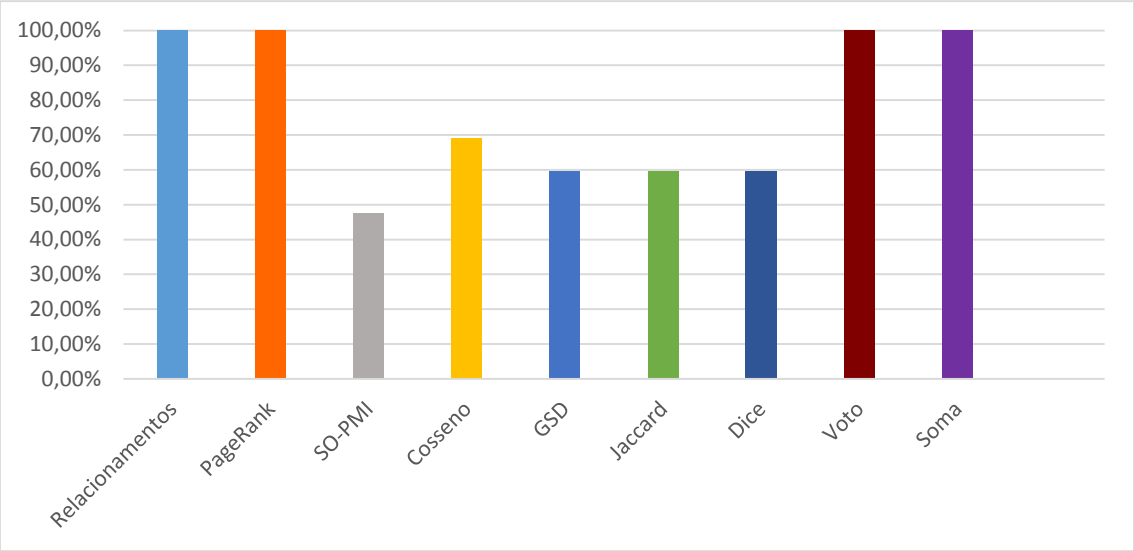

Figura 14 – Percentuais de cobertura das classificações dos termos presentes no TeP 2.0, por cada algoritmo avaliado.

Além disso, é importante dizer que, apesar de suas características, a base de *tweets* demonstrou ser capaz de alcançar resultados satisfatórios, principalmente se considerarmos que seu desempenho ficou no máximo 6% pior que o melhor desempenho dos algoritmos baseados no *corpus*, obtidos com a base do CETENFolha, como podemos ver na Figura 15 abaixo. Isso demonstra que textos extraídos do Twitter podem ser utilizados para este fim e que essa perda de desempenho é justificada com o ganho de informações obtidos ao se conseguir trabalhar e extrair termos de textos pouquíssimo estruturados, porém com altíssima influência na vida das pessoas atualmente. Detalhes sobre os resultados apresentados na Figura 15 são apresentados no Apêndice I deste trabalho.

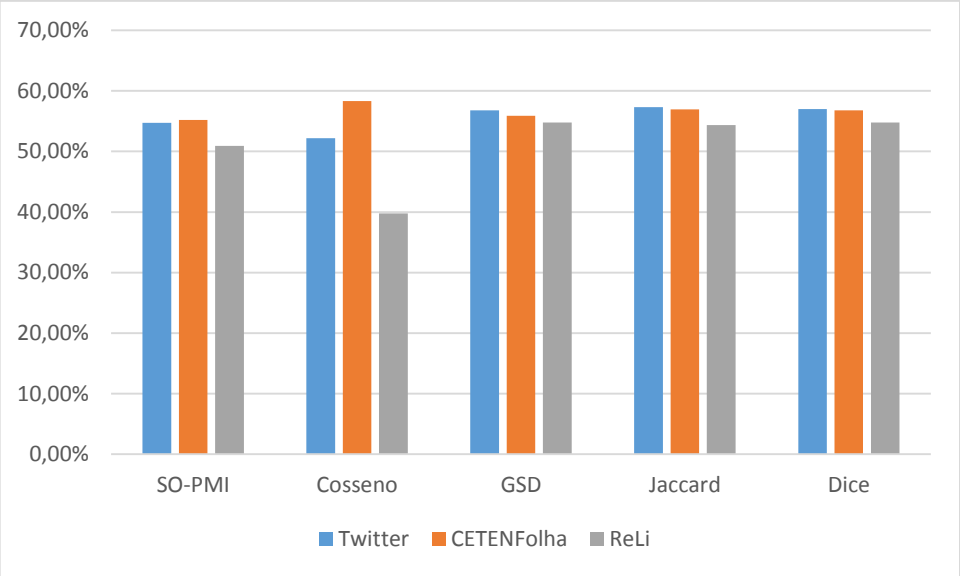

Figura 155 – Comparação entre os desempenhos de cada base, de acordo com o algoritmo de similaridade sintagmática utilizado.

Por fim, apresentamos na Tabela 25 abaixo alguns termos extraídos do *corpus* e classificados como positivos. Na Tabela 26, serão apresentados termos classificados como negativos. Essas classificações foram obtidas através da técnica de combinação por voto.

Tabela 25 – Exemplo de termos positivos extraídos do *corpus*.

Termo	Orientação
Irmão	Positivo
Músicas	Positivo
Fake	Positivo
Brasileiro	Positivo
Super	Positivo
Top	Positivo
Inteira	Positivo
Legalzinha	Positivo
Mãe	Positivo
Bombástica	Positivo

Tabela 26 – Exemplo de termos negativos extraídos do *corpus*.

Termo	Orientação
RS	Negativo
Fisiológica	Negativo
Ecaa	Negativo
Adolescente	Negativo
Fofocar	Negativo
Chatinha	Negativo
SOS	Negativo
Turbinada	Negativo
Zicas	Negativo
Monstro	Negativo

6 Conclusão

Este trabalho apresentou uma estratégia para a criação de um dicionário de polaridades em português do Brasil. Através de diversos experimentos foi possível identificar algoritmos capazes de cumprir esta tarefa com sucesso, permitindo a criação deste recurso de extrema importância para a área de Análise de Sentimento e ainda inexistente no nosso idioma. A proposta aqui apresentada também permite incluir no dicionário palavras extraídas dos documentos analisados, agregando conhecimento do domínio ao léxico gerado.

Esse trabalho apresenta ainda as seguintes contribuições originais:

1. Criação de um dicionário de polaridades em português do Brasil, contendo termos extraídos do *corpus* analisado;
2. Comparação do desempenho de diferentes métodos de cálculo de polaridades;
3. Implementação, para o português do Brasil, de estratégias já utilizadas para outros idiomas;
4. Aplicação de tais estratégias em um *corpus* composto por documentos oriundos do Twitter e escritos em português do Brasil.
5. Proposta do uso de estratégias compostas, que resultaram em desempenhos superiores aos de algumas das estratégias isoladas.

Como foi visto, a estratégia aqui apresentada parte de dois dicionários de origem, um contendo informações sobre as relações de sinonímia e antonímia entre termos em português do Brasil, o TeP 2.0, e o outro contendo informações sobre a polaridade de termos em português de Portugal, o SentiLex. Como o segundo foi criado com base no primeiro, foi possível extrair desses dicionários um conjunto de termos classificados manualmente, escritos em português do Brasil e com as informações sobre seus sinônimos e antônimos. Esse conjunto de 3636 termos foram utilizados como base para os experimentos de validação cruzada realizados para cada um dos classificadores analisados.

No total, foram 7 algoritmos avaliados, gerando 7 polaridades distintas para cada termo. Os dois primeiros algoritmos baseavam-se nas relações entre os termos para calcular seu repasse de polaridades. O de Análise de Relações atingiu acurácia de 78% ao calcular as polaridades dos termos sinônimos e antônimos considerando o decaimento do

valor repassado à medida que a distância entre os termos em um grafo aumenta. Já o Page Rank alcançou acurácia de quase 74% ao ser utilizado para calcular o quanto da polaridade de cada termo seria repassado para o seguinte ligado a ele.

Os outros métodos baseavam-se em medidas capazes de analisar a similaridade existente entres os termos. O primeiro deles, o SO-PMI, atingiu acurácia de até 55% ao analisar a coocorrência dos termos para calcular a probabilidade de os dois fazerem parte de uma mesma orientação. Esse algoritmo demonstrou ser altamente dependente do *corpus* utilizado, melhorando seu desempenho ao ser testado em bases com maior quantidade de informação sobre o comportamento dos termos no texto. O segundo deles, a Similaridade por Cosseno, inicialmente considerado uma contribuição inédita ao ser aplicado para esta atividade, demonstrou bom desempenho ao ser aplicada para o propósito de cálculo de polaridades, alcançando acurácia de até 58%. Além desse, também foi avaliado o algoritmo Google Similarity Distance, com acurácia de quase 56%, e cálculos clássicos de similaridade, como os Coeficientes de Jaccard e de Dice, que obtiveram resultados muito similares e chegaram a uma acurácia de quase 57%.

Esses últimos algoritmos, assim como a estratégia de extração de termos do *corpus* através da aplicação de regras com base na estrutura sintática dos textos, foram avaliados em cima de um *corpus* criado a partir da coleta de *tweets* sobre o Big Brother Brasil. A intenção era explorar um assunto bastante comentado e com grande carga subjetiva para avaliar o desempenho da estratégia aqui proposta ao trabalhar com textos altamente não estruturados, como os gerados através do Twitter. Devido à estrutura dos textos oriundos do Twitter e também para avaliar o impacto dessa estrutura nos resultados, outras bases foram analisadas e o Twitter demonstrou desempenho inferior de apenas 6%, mostrando que é possível utilizá-lo para esta atividade e usufruir dos benefícios de conseguir incorporar termos específicos desse domínio ao dicionário.

Como última estratégia, as diferentes polaridades geradas ainda foram combinadas por voto e por soma, chegando a alcançar acurácia de 66%, elevando em até 11% alguns dos resultados obtidos previamente.

Alguns pontos não puderam ser explorados neste trabalho e se apresentam como oportunidades futuras. Ao longo de toda a pesquisa realizada, ficou clara a falta de estratégias que consigam identificar novos sinônimos e antônimos semânticos a partir da

análise de textos. Como pudemos observar, essas relações são de extrema importância para conseguirmos desvendar a polaridade de termos ainda desconhecidos e continuar este trabalho por este caminho seria interessante por permitir incluir esta informação no dicionário, também em relação aos termos extraídos do *corpus*.

Além disso, neste trabalho optamos por avaliar o SO-PMI dentro do *corpus* analisado para extrairmos informações sobre o comportamento dos textos o mais contextualizadas possível. No entanto, seria interessante avaliar este mesmo algoritmo utilizando informações oriundas de mecanismos de busca da Internet. Essas ferramentas possuem vasta informação sobre a coocorrência de termos e podem auxiliar na melhoria dos resultados aqui obtidos.

Outra oportunidade futura seria também avaliar o ganho na classificação de textos em português do Brasil ao fazer uso do dicionário aqui gerado, tantos em textos vindos do Twitter como em qualquer tipo de texto.

Por fim, podemos concluir que a estratégia de criação de dicionários de polaridades aqui apresentada pode, de fato, ser aplicada à geração de léxicos em português do Brasil. Esta constatação é bastante importante para as iniciativas da área de Análise de Sentimentos por tornar possível que elas tenham acesso a este importante recurso, antes inexistente para o nosso idioma. Além disso, todo o processo aqui explicitado pode ser reproduzido para qualquer domínio de interesse, permitindo que cada iniciativa consiga agregar conhecimento contextualizado sobre os termos que compõem seus dicionários, contribuindo para a melhoria dos resultados obtidos pela Análise de Sentimento no nosso idioma.

Referências Bibliográficas

1. 140 CHARACTERS. How Twitter Was Born?, 2009. Disponível em <http://www.140characters.com/2009/01/30/how-twitter-was-born/>, acesso em 01/10/2012.

2. ACHREKAR, H.; GANDHE, A.; LAZARUS, R.; YU, S-H.; LIU, B. (2011) "Predicting Flu Trends using Twitter Data". The First International Workshop on Cyber-Physical Networking Systems, pp. 702-207.

3. BENEVENUTO, F.; MAGNO, G.; RODRIGUES, T.; ALMEIDA, V. (2010) "Detecting Spammers on Twitter". Proceedings of the Seventh Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference.

4. BOLLEN, J.; PEPE, A.; MAO, H. (2011) "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.". Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

5. BREW, A.; GREENE, D.; ARCHAMBAULT, D.; CUNNINGHAM, P. (2011) "Deriving Insights from National Happiness Indices". IEEE 11th International Conference On Data Mining Workshops (ICDMW), pp. 53 –60.

6. BRUCE, R.F.; WIEBE, J.M. (1999) "Recognizing subjectivity: a case study in manual tagging". Natural Language Engineering, v. 5, n. 2, p. 187-205.

7. CARVALHO, P.; SARMENTO, L.; SILVA, M. J.; OLIVEIRA, E. (2009) "Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-)". Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, pp. 53-56.

8. CASTILLO, C.; MENDOZA, M.; POBLETE, B. (2011) "Information credibility on Twitter". Proceedings of the 20th International Conference on World Wide Web, pp. 675-684.

9. CATALDI, M.; DI CARO, L.; SCHIFANELLA, C. (2010) "Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation". Proceedings of the Tenth International Workshop on Multimedia Data Mining.

10. CHA, M.; HADDADI, H.; BENEVENUTO, F.; GUMMADI, K.P. (2010) "Measuring User Influence in Twitter: The Million Follower Fallacy". Proceedings of the 4th International AAAI Conference on Weblogs and Social Media.

11. CHU, Z.; GIANVECCHIO, S.; WANG, H.; JAJODIA, S. (2010) "Who is tweeting on Twitter: human, bot, or cyborg?". Proceedings of the 26th Annual Computer Security Applications Conference, pp. 21-30.

12. CILIBRASI, R. L., VITANYI, P. M. (2007) "The Google Similarity Distance" IEEE Transactions on Knowledge and Data Engineering, pp. 370-383.

13. CONOVER, M.D.; GONÇALVES, B.; RATKIEWICZ, J.; FLAMMINI, A.; MENCZER, F. (2011) "Predicting the Political Alignment of Twitter Users".

Proceedings of 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, pp. 192-199.

14. CULOTTA, A. (2010) "Towards detecting influenza epidemics by analyzing Twitter messages". 1st Workshop on Social Media Analytics.

15. DAVIDOV, D.; TSUR, O.; and RAPPOPORT, A. (2010). "Enhanced sentiment learning using Twitter hashtags and smileys". Proceedings of the 23rd International Conference on Computational Linguistics, pp. 241–249.

16. DIAS-DA-SILVA, B. C. (2010). "Brazilian Portuguese WordNet: A Computational Linguistic Exercise of Encoding Bilingual Relational Lexicons". International Journal of Computational Linguistics and Applications, New Delhi, v.1, n. 1-2, p.137 – 150.

17. DIAS-DA-SILVA, B.C.; MORAES, H.R. (2003). "A construção de um thesaurus eletrônico para o português do Brasil". ALFA, Vol. 47, N. 2, pp. 101-115.

18. DUNLAP, J.C.; LOWENTHAL, P.R. (2009) "Tweeting the night away: Using Twitter to enhance social presence". Journal of Information Systems Education Special Issue, Impacts of Web 2.0 and Virtual World Technologies on IS Education, 20(2).

19. EARLE, P.S.; BOWDEN, D.C.; GUY, M. (2011) "Twitter earthquake detection: earthquake monitoring in a social world". Annals of Geophysics, Vol. 54, No. 6.

20. EBNER, M.; SCHIEFNER, M. (2008) "Microblogging – more than fun?". Proceedings of IADIS Mobile Learning Conference, pp. 155-159.

21. ESULI, A.; SEBASTIANI, F. (2005) "Determining the Semantic Orientation of Terms through Gloss Classification". Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 617-624.

22. ESULI, A.; SEBASTIANI, F. (2006) "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining". Proceedings of Language Resources and Evaluation, Vol. 6, pp. 417-422.

23. ESULI, A.; SEBASTIANI, F. (2007). "PageRanking Wordnet Synsets: An Application to Opinion Mining". Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 424–431.

24. FREITAS, C.; MOTTA, E.; MILIDIÚ, R.; CÉSAR, J. (2012). "Vampiro que brilha rá! Desafios na anotação de opinião em um corpus de resenhas de livros". ENCONTRO DE LINGUÍSTICA DE CORPUS, Vol. 11.

25. GATE. Developing Language Processing Components with GATE Version 8 (a User Guide), 2014. Disponível em <http://gate.ac.uk/sale/tao/split.html>, acesso em 25/06/2014.

26. GODBOLE, N.; SRINIVASAIAH, M.; SKIENA, S. (2007) "Large-Scale Sentiment Analysis for News and Blogs". International Conference on Weblogs and Social Media, Vol. 7.

27. GOLDEN, P. Write here, write now, 2011. Disponível em <http://www.research-live.com/features/write-here-write-now/4005303.article>, acesso em 08/10/2012.

28. HATZIVASSILOGLOU, V.; MCKEOWN, K. R. (1997) “Predicting the Semantic Orientation of Adjectives”. Proceedings of Annual Meeting of the Association for Computational Linguistics.

29. HATZIVASSILOGLOU, V.; WIEBE, J.M. (2000) “Effects of adjective orientation and gradability on sentence subjectivity”. Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics. p. 299-305.

30. HERRING, S.C.; SCHEIDT, L.A.; BONUS, S.; WRIGHT, E. (2004) “Bridging the Gap: A Genre Analysis of Weblogs”. Proceedings of the 37th Hawaii International Conference on System Sciences, Track 4, Vol. 4.

31. HONEYCUTT, C.; HERRING, S.C. (2009) “Beyond Microblogging: Conversation and Collaboration via Twitter”. Proceedings of the 42nd Hawaii International Conference on System Sciences, pp. 1-10.

32. HU, M.; LIU, B. (2004) “Mining and Summarizing Customer Reviews”. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168-177.

33. HU, X.; TANG, L.; TANG, J.; LIU, H. (2013). “Exploiting social relations for sentiment analysis in microblogging”. Proceedings of the 6th ACM International Conference on Web Search and Data Mining, pp. 537-546.

34. HUBERMAN, B.A.; ROMERO, D.M.; WU, F. (2008) “Social networks that matter: Twitter under the microscope”. Available at SSRN: http://ssrn.com/abstract=1313405 or http://dx.doi.org/10.2139/ssrn.1313405.

35. JAVA, A.; SONG, X.; FININ, T.; TSENG, B. (2007) “Why we twitter: understanding microblogging usage and communities”. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pp. 56-65.

36. JIANG, L.; YU, M.; ZHOU, M.; LIU, X.; Zhao, T. (2011) “Target-dependent Twitter sentiment classification”, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, pp. 151–160.

37. JUNG. Overview, 2010. Disponível em <http://jung.sourceforge.net/>, acesso em 25/06/2014.

38. KAMPS, J.; MARX, M.; MOKKEN, R. J.; RIJKE, M. (2004) “Using WordNet to measure semantic orientation of adjectives”. Proceedings of Language Resources and Evaluation Conference.

39. KIM, H. D.; ZHAI, C. (2009) “Generating Comparative Summaries of Contradictory Opinions in Text”. Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 385-394.

40. KIM, S. M.; HOVY, E. (2004) “Determining the sentiment of opinions”. Proceedings of the 20th International Conference on Computational Linguistics, pp. 1367.

41. KRISHNAMURTHY, B.; GILL, P.; ARLITT, M. (2008) “A few chirps about twitter”. Proceedings of the 1st Workshop on Online Social Networks, pp. 19-24.

42. KWAK, H.; LEE, C.; PARK, H.; MOON, S. (2010) “What is Twitter, a social network or a news media?”. Proceedings of the 19th International Conference on World Wide Web, pp. 591–600.

43. LAFFERTY, J.; McCALLUM, A.; PEREIRA, F. C. (2001). “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”.

44. LI, X.-L.; LEI, Z.; BING, L.; SEE-KIONG, N. (2010) “Distributional Similarity vs. PU Learning for Entity Set Expansion”. Proceedings of Annual Meeting of the Association for Computational Linguistics.

45. LI, Y.-M.; LI, T.-Y. (2011) “Deriving Marketing Intelligence over Microblogs”. Proceedings of 44th Hawaii International Conference on System Sciences (HICSS), pp. 1 –10.

46. LINGUATECA. CETENFolha. 2002. Disponível em < http://www.linguateca.pt/cetenfolha/index_info.html>, acesso em 25/06/2014.

47. LIU, B. (2010) “Sentiment Analysis and Subjectivity”. Handbook of Natural Language Processing, 2nd Edition.

48. LIU, B. (2012) “Sentiment analysis and opinion mining”. Synthesis Lectures on Human Language Technologies, v. 5, n. 1, pp. 1-167.

49. LUCENE. Apache Lucene Core, 2012. Disponível em <http://lucene.apache.org/core/>, acesso em 25/06/2014.

50. MALONE, T. W.; LAUBACHER, R.; DELLAROCAS C. (2009) “Harnessing Crowds: Mapping the Genome of Collective Intelligence”. MIT Center for Collective Intelligence, Working Paper No. 2009-001.

51. MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. (2008) “Introduction to Information Retrieval”, Vol. 1. Cambridge: Cambridge University Press.

52. MANNING, C. D.; SCHÜTZE, H. (1999). “Foundations of Statistical Natural Language Processing”. MIT Press.

53. MATHIOUDAKIS, M.; KOUDAS, N. (2010). “TwitterMonitor: Trend Detection over the Twitter Stream”. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 1155-1158.

54. MAZIERO, E.G.; PARDO, T.A.S.; DI FELIPPO, A.; DIAS-DA-SILVA, B.C. (2008). “A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil”. VI WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TIL), pp. 390-392.

55. METAXAS, P.T.; MUSTAFARAJ, E.; GAYO-AVELLO, D. (2011) “How (Not) To Predict Elections”. Proceedings of 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, pp. 165-171.

56. MISCHAUD, E. Twitter: Expressions of the whole self. An investigation into user appropriation of a web-based communications platform, 2007. Disponível em <

http://www2.lse.ac.uk/media@lse/research/mediaWorkingPapers/MScDissertationSeries/Mishaud_Final.pdf>, acesso em: 21/01/2013.

57. MISSEN, M.; BOUGHANEM, M.; CABANAC, G. (2013) “Opinion mining: reviewed from word to document level”. Social Network Analysis and Mining, v. 3, n. 1, pp. 107-125.

58. NADEAU, D.; SEKINE, S. (2007) “A survey of named entity recognition and classification”. Lingvisticae Investigationes, v. 30, n. 1, p. 3-26.

59. NAAMAN, M.; BOASE, J.; Lai, C.-H. (2010) “Is it all About Me? User Content in Social Awareness Streams”. Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work.

60. NEO4J. What is Neo4J?, 2014. Disponível em <http://www.neo4j.org/learn/neo4j>, acesso em 25/06/2014.

61. NIELSENWIRE. Buzz in the Blogosphere: Millions More Bloggers and Blog Readers, 2012. Disponível em <http://blog.nielsen.com/nielsenwire/online_mobile/buzz-in-the-blogosphere-millions-more-bloggers-and-blog-readers/>, acesso em 16/11/2012.

62. PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. (1999) “The PageRank Citation Ranking: Bringing Order to the Web”.

63. PAK, A.; PAROUBEK, P. (2010). “Twitter as a corpus for sentiment analysis and opinion mining”. Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC’10).

64. PANG, B.; LEE, L.; VAITHYANATHAN, S. (2002) "Thumbs up?: sentiment classification using machine learning techniques". Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Vol. 10.

65. PANG, B.; LEE, L. (2004) “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”. Proceedings of the 42nd annual meeting on Association for Computational Linguistics, pp. 271.

66. PANG, B.; LEE, L. (2008) “Opinion Mining and Sentiment Analysis”. Foundations and Trends in Information Retrieval, v. 2, n. 1-2, pp. 1-135.

67. PETROVIC, S.; OSBORNE, M.; LAVRENKO, V. (2010) “Streaming First Story Detection with application to Twitter”. Proceeding of HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 181-189.

68. PHUVIPADAWAT, S.; MURATA, T. (2010) “Breaking News Detection and Tracking in Twitter”. Proceedings of 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 120-123.

69. QIU, G.; LIU, B.; BU, J.; CHEN, C. (2009) “Expanding Domain Sentiment Lexicon Through Double Propagation”. Proceedings of the 21st International Jont Conference on Artificial Intelligence, pp. 1199-1204.

70. RILOFF, E.; WIEBE, J.; WILSON, T. (2003) “Learning subjective nouns using extraction pattern bootstrapping”. Proceedings of the 7th conference on Natural Language Learning, vol. 4, pp. 25-32.

71. RILOFF, E.; WIEBE, J.; PHILIPS, W. (2005) "Exploiting subjectivity classification to improve information extraction." Proceedings of the National Conference On Artificial Intelligence. Vol. 20. No. 3.

72. SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. (2010) “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors”. Proceedings of the 19th international conference on World Wide Web, pp. 851-860.

73. SANTOS, A. P.; RAMOS, C.; MARQUES, N. C. (2011) “Determining the polarity of words through a common online dictionary”. Progress in Artificial Intelligence, pp. 649-663. Springer Berlin Heidelberg.

74. SARAWAGI, S. (2008). “Information extraction”. Foundations and Trends in Databases, Vol. 1, No. 3, pp. 261–377.

75. SEMIOCAST. Twitter reaches half a billion accounts. More than 140 millions in the U.S., 2012. Disponível em <http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US>, acesso em 08/10/2012.

76. SILVA, M. J.; CARVALHO, P.; COSTA, C.; SARMENTO, L. (2010) “Automatic expansion of a social judgment lexicon for sentiment analysis”. Technical Report TR 1008 University of Lisbon Faculty of Sciences LASIGE.

77. SOUZA, M.; VIEIRA, R.; BUSETTI, D.; CHICHMAN, R.; ALVES, I. M. (2011). “Construction of a Portuguese Opinion Lexicon from multiple resources”. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pp. 59-66.

78. STELZNER, M. A. 2012 Social Media Marketing Industry Report, 2012. Disponível em < http://www.socialmediaexaminer.com/SocialMediaMarketingIndustryReport2012.pdf>, acesso em 08/10/2012.

79. SLOMAN, A.; CHRISLEY, R.; SCHEUTZ, M. (2005) “The architectural basis of affective states and processes”. Who Needs Emotions?: The Brain Meets the Machine, v. 3, pp. 203–244.

80. TURNEY, P. D. (2002). “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 417-424.

81. TURNEY, P. D.; LITTMAN, M. L. (2003). “Measuring Praise and Criticism: Inference of Semantic Orientation from Association”. ACM Transactions on Information Systems (TOIS), v. 21, n. 4, pp. 315-346.

82. TWITTER. About Twitter, 2012. Disponível em <https://twitter.com/about>, acesso em 01/10/2012.

83. TWITTER. Twitter 101: Como começar a usar o Twitter?, 2013a. Disponível em < https://support.twitter.com/groups/31-twitter-basics/topics/104-welcome-to-twitter-support/articles/262253-twitter-101-como-comecar-a-usar-o-twitter>, acesso em 24/01/2013.

84. TWITTER. The Twitter Glossary, 2013b. Disponível em < https://support.twitter.com/articles/166337-the-twitter-glossary>, acesso em 24/01/2013.

85. VIEWEG, S.; HUGHES, A.L.; STARBIRD, K.; PALEN, L. (2010) "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness". Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1079-1088.

86. XIA, R; ZONG, C.; LI, S. (2011). "Ensemble of feature sets and classification algorithms for sentiment classification." Information Sciences, Vol. 181, No. 6, pp. 1138-1152.

87. WIEBE, J.M.; BRUCE, R.F.; O'HARA, T.P. (1999) "Development and use of a gold-standard data set for subjectivity classifications". Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, p. 246 – 253.

88. WIEBE, J.; WILSON, T.; CARDIE, C. (2006) "Annotating Expressions of Opinions and Emotions in Language". Language Resources and Evaluation, v. 39, n. 2-3, pp. 165 – 210.

89. WEKA. Weka 3: Data Mining Software in Java, 2014. Disponível em <http://www.cs.waikato.ac.nz/ml/weka/>, acesso em 24/08/2014.

90. WORDNET. What is Wordnet?, 2013. Disponível em < http://wordnet.princeton.edu/>, acesso em 27/05/2014.

91. XU, Z.; ZHANG, Y.; WU, Y.; YANG, Q. (2012) "Modeling user posting behavior on social media". Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 545-554.

92. YOUNUS, A.; QURESHI, M.A.; ASAR, F.F.; AZAM, M.; SAEED, M.; TOUHEED, N. (2011) "What do the Average Twitterers Say: a Twitter Model for Public Opinion Analysis in the Face of Major Political Events". Proceedings of 2011 International Conference on Advances in Social Networks Analysis and Mining, pp. 618-623.

93. ZHANG, K.; CHENG, Y.; XIE, Y.; HONBO, D.; AGRAWAL, A.; PALSETIA, D.; LEE, K.; LIAO, W.; CHOUDHARY, A. (2011). "SES: Sentiment Elicitation System for Social Media Data". Proceedings of 11th International Conference on Data Mining Workhops (ICDMW), pp. 129 – 136.

94. ZHAO, D.; ROSSON, M. B. (2009) "How and Why People Twitter: The Role that Micro-blogging Plays in Informal Communication at Work". Proceedings of the ACM 2009 International Conference on Supporting Group Work, pp. 243-252.

Apêndice I – Valores numéricos para todos os resultados alcançados

Este apêndice apresenta todos os resultados obtidos para todas as bases utilizadas ao avaliarmos os algoritmos de cálculo de polaridades. Um resumo dos melhores destes resultados é apresentado no capítulo 5, junto aos resultados dos algoritmos independentes do *corpus*, onde uma análise sobre os mesmos também é realizada.

1. Resultados da Análise de Relações

Tabela 27 – Valores médios do algoritmo Análise de Relações para cada orientação.

Acurácia: 78,0265%			
	Positivo	**Negativo**	**Neutro**
Precisão	69,4390%	**88,3311%**	28,5556%
Revocação	**87,0628%**	85,8005%	16,5972%

2. Resultados do Page Rank

Tabela 28 – Valores médios do algoritmo Page Rank para cada orientação.

Acurácia: 73,8027%			
	Positivo	**Negativo**	**Neutro**
Precisão	54,0676%	**89,9627%**	0%
Revocação	**92,44%**	80,0128%	0%

3. Resultados do SO-PMI

Tabela 29 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* de *tweets*.

Acurácia: 50,7232%			
	Positivo	**Negativo**	**Neutro**
Precisão	38,2382%	**59,9323%**	0%
Revocação	49,9988%	**61,7211%**	0%

Tabela 30 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=7, quando usado o *corpus* de *tweets*.

Acurácia: 51,8655%			
	Positivo	**Negativo**	**Neutro**
Precisão	37,9818%	**63,8934%**	0%
Revocação	56,4514%	**61,7211%**	0%

Tabela 31 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=10, quando usado o *corpus* de *tweets*.

Acurácia: 54,7481%			
	Positivo	**Negativo**	**Neutro**
Precisão	41,1867%	**62,2055%**	0%
Revocação	46,6424%	**70,8422%**	0%

Tabela 32 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=15, quando usado o *corpus* de *tweets*.

Acurácia: 50,7775%			
	Positivo	**Negativo**	**Neutro**
Precisão	36,3055%	**61,0530%**	0%
Revocação	48,6557%	**62,5553%**	0%

Tabela 33 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=20, quando usado o *corpus* de *tweets*.

Acurácia: 51,4899%			
	Positivo	**Negativo**	**Neutro**
Precisão	35,7633%	**60,3872%**	0%
Revocação	42,4899%	**66,9485%**	0%

Tabela 34 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* ReLi.

Acurácia: 49,7845%			
	Positivo	**Negativo**	**Neutro**
Precisão	30,7496%	**57,1600%**	0%
Revocação	28,5898%	**71,4989%**	0%

Tabela 35 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=7, quando usado o *corpus* ReLi.

Acurácia: 50,9223%			
	Positivo	**Negativo**	**Neutro**
Precisão	29,5241%	**58,2599%**	0%
Revocação	30,9755%	**72,2363%**	0%

Tabela 36 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=10, quando usado o *corpus* ReLi.

Acurácia: 50,8582%			
	Positivo	**Negativo**	**Neutro**
Precisão	34,6072%	**51,9759%**	0%
Revocação	18,9302%	**80,1465%**	0%

Tabela 37 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=15, quando usado o *corpus* ReLi.

Acurácia: 49,8263%			
	Positivo	**Negativo**	**Neutro**
Precisão	30,6216%	**57,4646%**	0%
Revocação	29,3142%	**71,1823%**	0%

Tabela 38 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=20, quando usado o *corpus* ReLi.

Acurácia: 50,4761%			
	Positivo	**Negativo**	**Neutro**
Precisão	31,1608%	**57,3214%**	0%
Revocação	26,9258%	**73,5841%**	0%

Tabela 39 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* CETENFolha.

Acurácia: 55,2012%			
	Positivo	**Negativo**	**Neutro**
Precisão	40,0040%	**61,2829%**	0%
Revocação	38,4753%	**76,1243%**	0%

Tabela 40 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=7, quando usado o *corpus* CETENFolha.

Acurácia: 45,7922%			
	Positivo	**Negativo**	**Neutro**
Precisão	27,8225%	**58,7543%**	0%
Revocação	41,1829%	**56,9548%**	0%

Tabela 41 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=10, quando usado o *corpus* CETENFolha.

Acurácia: 50,4623%			
	Positivo	**Negativo**	**Neutro**
Precisão	29,0448%	**59,8631%**	0%
Revocação	31,3088%	**69,7352%**	0%

Tabela 42 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=15, quando usado o *corpus* CETENFolha.

Acurácia: 52,4566%			
	Positivo	**Negativo**	**Neutro**
Precisão	33,0449%	**48,1195%**	0%
Revocação	28,8084%	**74,1244%**	0%

Tabela 43 – Valores médios do algoritmo SO-PMI para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=20, quando usado o *corpus* CETENFolha.

Acurácia: 51,3960%			
	Positivo	**Negativo**	**Neutro**
Precisão	29,6352%	**54,2981%**	0%
Revocação	27,2432%	**71,1553%**	0%

4. Resultados da Similaridade por Cosseno

Tabela 44 – Valores médios do algoritmo Similaridade por Cosseno para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* de *tweets*.

Acurácia: 52,2068%			
	Positivo	**Negativo**	**Neutro**
Precisão	50,1905%	**58,4009%**	10,9973%
Revocação	3,6570%	**84,9110%**	11,9093%

Tabela 45 – Valores médios do algoritmo Similaridade por Cosseno para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* ReLi.

Acurácia: 39,7792%			
	Positivo	**Negativo**	**Neutro**
Precisão	38,7360%	**53,6208%**	10,3894%
Revocação	16,8004%	**56,7904%**	20,7274%

Tabela 46 – Valores médios do algoritmo Similaridade por Cosseno para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* CETENFolha.

Acurácia: 58,2991%			
	Positivo	**Negativo**	**Neutro**
Precisão	34,2857%	**60,1106%**	4,2063%
Revocação	0,6793%	**95,1385%**	4,2662%

5. Resultados da *Google Similarity Distance*

Tabela 47 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* de *tweets*.

Acurácia: 50,8397%			
	Positivo	**Negativo**	**Neutro**
Precisão	34,0008%	**64,0987%**	0%
Revocação	52,4005%	**60,5609%**	0%

Tabela 48 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=7, quando usado o *corpus* de *tweets*.

Acurácia: 55,8653%			
	Positivo	**Negativo**	**Neutro**
Precisão	36,1729%	**61,8946%**	0%
Revocação	30,1360%	**79,4067%**	0%

Tabela 49 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=10, quando usado o *corpus* de *tweets*.

Acurácia: 56,7648%			
	Positivo	**Negativo**	**Neutro**
Precisão	37,0749%	**61,5384%**	0%
Revocação	25,4733%	**83,2015%**	0%

Tabela 50 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=15, quando usado o *corpus* de *tweets*.

Acurácia: 55,6869%			
	Positivo	**Negativo**	**Neutro**
Precisão	34,5351%	**60,9824%**	0%
Revocação	24,3825%	**81,7476%**	0%

Tabela 51 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=20, quando usado o *corpus* de *tweets*.

Acurácia: 55,4331%			
	Positivo	**Negativo**	**Neutro**
Precisão	34,9558%	**61,4717%**	0%
Revocação	28,7679%	**79,1855%**	0%

Tabela 52 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* ReLi.

Acurácia: 54,7653%			
	Positivo	**Negativo**	**Neutro**
Precisão	36,5781%	**63,7117%**	0%
Revocação	42,7275%	**71,3196%**	0%

Tabela 53 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=7, quando usado o *corpus* ReLi.

Acurácia: 52,8656%			
	Positivo	**Negativo**	**Neutro**
Precisão	30,7241%	**59,9249%**	0%
Revocação	26,3535%	**76,0235%**	0%

Tabela 54 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=10, quando usado o *corpus* ReLi.

Acurácia: 52,9594%			
	Positivo	**Negativo**	**Neutro**
Precisão	30,8319%	**59,8509%**	0%
Revocação	25,8807%	**76,4110%**	0%

Tabela 55 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=15, quando usado o *corpus* ReLi.

Acurácia: 51,3719%			
	Positivo	**Negativo**	**Neutro**
Precisão	30,1432%	**59,4656%**	0%
Revocação	29,6719%	**71,9381%**	0%

Tabela 56 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=20, quando usado o *corpus* ReLi.

Acurácia: 51,1126%			
	Positivo	**Negativo**	**Neutro**
Precisão	31,6066%	**60,5055%**	0%
Revocação	36,4964%	**68,1943%**	0%

Tabela 57 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* CETENFolha.

Acurácia: 47,3422%			
	Positivo	**Negativo**	**Neutro**
Precisão	31,4372%	**63,8887%**	0%
Revocação	56,1791%	**52,7594%**	0%

Tabela 58 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=7, quando usado o *corpus* CETENFolha.

Acurácia: 54,0848%			
	Positivo	**Negativo**	**Neutro**
Precisão	36,4020%	**60,9070%**	0%
Revocação	25,5931%	**79,0245%**	0%

Tabela 59 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=10, quando usado o *corpus* CETENFolha.

Acurácia: 55,8676%			
	Positivo	**Negativo**	**Neutro**
Precisão	32,9554%	**60,7649%**	0%
Revocação	20,3141%	**84,5633%**	0%

Tabela 60 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=15, quando usado o *corpus* CETENFolha.

Acurácia: 52,6564%			
	Positivo	**Negativo**	**Neutro**
Precisão	29,5207%	**60,6871%**	0%
Revocação	26,3985%	**76,0817%**	0%

Tabela 61 – Valores médios do algoritmo NGD para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=20, quando usado o *corpus* CETENFolha.

Acurácia: 51,1904%			
	Positivo	**Negativo**	**Neutro**
Precisão	29,9283%	**61,4196%**	0%
Revocação	34,4370%	**69,5887%**	0%

6. Resultados do Coeficiente de Dice

Tabela 62 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* de *tweets*.

Acurácia: 52,6242%			
	Positivo	**Negativo**	**Neutro**
Precisão	30,0709%	**56,2425%**	0%
Revocação	39,2068%	**69,8568%**	0%

Tabela 63 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=7, quando usado o *corpus* de *tweets*.

Acurácia: 56,5332%			
	Positivo	**Negativo**	**Neutro**
Precisão	39,1609%	**64,2592%**	0%
Revocação	41,7029%	**78,0241%**	0%

Tabela 64 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=10, quando usado o *corpus* de *tweets*.

Acurácia: 57,0083%			
	Positivo	**Negativo**	**Neutro**
Precisão	37,9729%	**62,5489%**	0%
Revocação	29,4443%	**81,6208%**	0%

Tabela 65 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=15, quando usado o *corpus* de *tweets*.

Acurácia: 55,1751%			
	Positivo	**Negativo**	**Neutro**
Precisão	34,2323%	**61,7798%**	0%
Revocação	28,6620%	**78,8434%**	0%

Tabela 66 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=20, quando usado o *corpus* de *tweets*.

Acurácia: 55,8789%			
	Positivo	**Negativo**	**Neutro**
Precisão	36,5316%	**62,6435%**	0%
Revocação	28,7679%	**84,5505%**	0%

Tabela 67 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* ReLi.

Acurácia: 54,7653%			
	Positivo	**Negativo**	**Neutro**
Precisão	35,5885%	**63,2662%**	0%
Revocação	41,9624%	**70,5678%**	0%

Tabela 68 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=7, quando usado o *corpus* ReLi.

Acurácia: 52,7242%			
	Positivo	**Negativo**	**Neutro**
Precisão	29,3554%	**59,5250%**	0%
Revocação	23,3002%	**77,1945%**	0%

Tabela 69 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=10, quando usado o *corpus* ReLi.

Acurácia: 52,8166%			
	Positivo	**Negativo**	**Neutro**
Precisão	29,0671%	**59,4396%**	0%
Revocação	22,2425%	**77,8512%**	0%

Tabela 70 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=15, quando usado o *corpus* ReLi.

Acurácia: 52,0568%			
	Positivo	**Negativo**	**Neutro**
Precisão	26,6936%	**59,4543%**	0%
Revocação	26,2008%	**74,7150%**	0%

Tabela 71 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=20, quando usado o *corpus* ReLi.

Acurácia: 51,7259%			
	Positivo	**Negativo**	**Neutro**
Precisão	30,3709%	**60,2119%**	0%
Revocação	30,6551%	**72,0719%**	0%

Tabela 72 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* CETENFolha.

Acurácia: 52,0405%			
	Positivo	**Negativo**	**Neutro**
Precisão	33,3625%	**66,0746%**	0%
Revocação	51,0834%	**62,9425%**	0%

Tabela 73 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=7, quando usado o *corpus* CETENFolha.

Acurácia: 52,3884%			
	Positivo	**Negativo**	**Neutro**
Precisão	24,1498%	**58,4671%**	0%
Revocação	19,9456%	**80,4347%**	0%

Tabela 74 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=10, quando usado o *corpus* CETENFolha.

Acurácia: 55,6709%			
	Positivo	**Negativo**	**Neutro**
Precisão	21,3220%	**59,0944%**	0%
Revocação	6,6612%	**89,7833%**	0%

Tabela 75 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=15, quando usado o *corpus* CETENFolha.

Acurácia: 56,7752%			
	Positivo	**Negativo**	**Neutro**
Precisão	23,3071%	**59,4653%**	0%
Revocação	5,9301%	**92,0069%**	0%

Tabela 76 – Valores médios do algoritmo Coeficiente de Dice para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=20, quando usado o *corpus* CETENFolha.

Acurácia: 55,7775%			
	Positivo	**Negativo**	**Neutro**
Precisão	25,4523%	**59,7428%**	0%
Revocação	10,5371%	**88,2805%**	0%

7. Resultados do Coeficiente de Jaccard

Tabela 77 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* de *tweets*.

Acurácia: 53,4129%			
	Positivo	**Negativo**	**Neutro**
Precisão	33,5605%	**62,3095%**	0%
Revocação	36,0204%	**72,7534%**	0%

Tabela 78 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=7, quando usado o *corpus* de *tweets*.

Acurácia: 56,8225%			
	Positivo	**Negativo**	**Neutro**
Precisão	39,1609%	**64,2592%**	0%
Revocação	39,4392%	**76,0772%**	0%

Tabela 79 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=10, quando usado o *corpus* de *tweets*.

Acurácia: 57,2976%			
	Positivo	**Negativo**	**Neutro**
Precisão	38,3240%	**62,4426%**	0%
Revocação	28,3623%	**82,6394%**	0%

Tabela 80 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=15, quando usado o *corpus* de *tweets*.

Acurácia: 55,4615%			
	Positivo	**Negativo**	**Neutro**
Precisão	34,3934%	**61,6816%**	0%
Revocação	27,5914%	**79,8543%**	0%

Tabela 81 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=20, quando usado o *corpus* de *tweets*.

Acurácia: 56,1976%			
	Positivo	**Negativo**	**Neutro**
Precisão	36,8731%	**62,5669%**	0%
Revocação	32,3235%	**78,8113%**	0%

Tabela 82 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* ReLi.

Acurácia: 54,3610%			
	Positivo	**Negativo**	**Neutro**
Precisão	35,8203%	**63,2087%**	0%
Revocação	40,9631%	**71,6208%**	0%

Tabela 83 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=7, quando usado o *corpus* ReLi.

Acurácia: 52,8559%			
	Positivo	**Negativo**	**Neutro**
Precisão	29,1931%	**59,4539%**	0%
Revocação	22,4052%	**77,8798%**	0%

Tabela 84 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=10, quando usado o *corpus* ReLi.

Acurácia: 52,9824%			
	Positivo	**Negativo**	**Neutro**
Precisão	29,0024%	**59,3956%**	0%
Revocação	21,4625%	**78,5372%**	0%

Tabela 85 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=15, quando usado o *corpus* ReLi.

Acurácia: 52,1880%			
	Positivo	**Negativo**	**Neutro**
Precisão	29,6095%	**59,3878%**	0%
Revocação	25,4208%	**75,3402%**	0%

Tabela 86 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_NEG extraídos do *corpus* e com N=20, quando usado o *corpus* ReLi.

Acurácia: 51,7920%			
	Positivo	**Negativo**	**Neutro**
Precisão	30,2374%	**60,1036%**	0%
Revocação	29,8751%	**72,5889%**	0%

Tabela 87 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_NEG originais, quando usado o *corpus* CETENFolha.

Acurácia: 52,0163%			
	Positivo	**Negativo**	**Neutro**
Precisão	33,8357%	**65,2492%**	0%
Revocação	50,5248%	**63,1961%**	0%

Tabela 88 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=7, quando usado o *corpus* CETENFolha.

Acurácia: 52,3427%			
	Positivo	**Negativo**	**Neutro**
Precisão	23,9435%	**58,2788%**	0%
Revocação	14,4578%	**80,6682%**	0%

Tabela 89 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=10, quando usado o *corpus* CETENFolha.

Acurácia: 55,7871%			
	Positivo	**Negativo**	**Neutro**
Precisão	20,9510%	**59,0104%**	0%
Revocação	6,1251%	**90,2590%**	0%

Tabela 90 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=15, quando usado o *corpus* CETENFolha.

Acurácia: 56,9737%			
	Positivo	**Negativo**	**Neutro**
Precisão	23,6772%	**59,4486%**	0%
Revocação	5,6068%	**92,5219%**	0%

Tabela 91 – Valores médios do algoritmo Coeficiente de Jaccard para cada orientação, utilizando G_POS e G_ NEG extraídos do *corpus* e com N=20, quando usado o *corpus* CETENFolha.

Acurácia: 55,7775%			
	Positivo	**Negativo**	**Neutro**
Precisão	25,4523%	**59,7428%**	0%
Revocação	10,5371%	**88,2805%**	0%