Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Uma Abordagem de Análise de Sentimentos Espaço-Temporal em Microtextos

André Luiz Firmino Alves

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Banco de Dados e Internet

Cláudio de Souza Baptista, Ph.D.
(Orientador)

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

A474a Alves, André Luiz Firmino.

Uma abordagem de análise de sentimentos espaço-temporal em microtextos / André Luiz Firmino Alves. – Campina Grande, 2014.

88 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2014.

"Orientação: Prof. Dr. Cláudio de Souza Baptista".

Referências.

1. Micro-Blogging. 2. Análise de Sentimentos. 3. Análise Espaço-Temporal. I. Baptista, Cláudio de Souza. II. Título.

CDU 004.771(043)

"UMA ABORDAGEM DE ANÁLISE DE SENTIMENTOS ESPAÇO-TEMPORAL EM MICROTEXTOS"

ANDRÉ LUIZ FIRMINO ALVES

DISSERTAÇÃO APROVADA EM 27/08/2014

Cláudio de Souza Baptista

CLÁUDIO DE SOUZA BAPTISTA, PhD., UFCG
Orientador(a)

CLAUDIO ELIZIO CALAZANS CAMPELO, PhD., UFCG
Examinador(a)

Fabio Gomes de Andrade

FABIO GOMES DE ANDRADE, D.Sc., IFPB
Examinador(a)

CAMPINA GRANDE - PB

Resumo

A proliferação dos meios de comunicação social na Web, tais como blogs, fóruns de discussões, sites de avaliação de produtos, microblogs e redes sociais, proporcionou um volume de dados opinativos armazenados em formato digital nunca visto na história da humanidade. Esta quantidade de dados, em sua grande maioria não estruturados, tem trazido vários desafios e oportunidades para a comunidade acadêmica e o mundo dos negócios, haja vista a necessidade de compreender, de forma automática, os sentimentos das pessoas a respeito de um produto, um serviço ou mesmo sobre pessoas ou fatos, para auxiliar no processo de tomada de decisão. Nos últimos anos, surgiram várias contribuições científicas para resolver problemas relacionados à análise de sentimentos. No entanto, poucas propostas consideram o fator espaço-temporal, isto é, a localização geográfica da fonte de informação ou da própria informação, bem como as possíveis mudanças de opinião ao longo do tempo. Os trabalhos que consideram o fator espacial tomam como base mensagens já geocodificadas, contudo, são poucas as fontes de informações que dispõem de mensagens georeferenciadas. Neste contexto, este trabalho propõe uma abordagem de análise de sentimentos que explora os fatores espaço-temporal para melhor sumarizar o sentimento detectado em uma grande quantidade de microtextos obtidos da Web. A abordagem utiliza técnicas de Recuperação da Informação Geográfica (GIR) e técnicas de Análise de Sentimentos para detectar localizações geográficas e a polaridade dos sentimentos através de evidências textuais contidas nos microtextos, oferecendo mecanismos de visualização espacial do sentimento em diversas regiões geográficas. A análise espaço-temporal possibilita visualizar mudanças de sentimento ocorridas em diversas regiões geográficas ao longo do período analisado.

Abstract

The dissemination of social communication means on the Web, such as blogs, discussion forums, product evaluation sites, microblogs and social networks, provides a never before seen volume of opinionative data in digital format. Not structured in its majority, this amount of data, has brought several challenges and opportunities for the academic community and the business world, considering the need for understanding, in an automatic form, people's sentiments concerning a product, a service or even other people or facts, in order to facilitate the decision making process. In the recent years, several scientific contributions to solve sentiment analysis related problems were suggested. However, only a few of them consider the spatial-temporal factor, which is the geographical location of the information source or even of the information itself, as well as the possible opinion changes throughout time. The works that consider the spatial factor often assume the messages are already geocoded. However, it could be a problem, since only a few information sources provide georeferenced messages. In this context, this work proposes a sentiment analysis approach which explores the spatial-temporal factor in order to better summarize the sentiments detected in a great amount of microtexts obtained from the Web. The approach uses Geographic Information Retrieval (GIR) and Sentiment Analysis techniques for the detection of geographic locations and sentiment polarity through textual evidences contained in the microtexts. The spatial-temporal analysis enables the visualization of sentiment changes which occurred in several geographic regions throughout the analyzed time period.

Agradecimentos

À Deus, por seu infinito amor, por me sustentar nos momentos em que não tinha mais força para continuar a jornada, pela inspiração e capacitação. Na verdade, sem Ele nada do que foi feito se fez. Obrigado Deus...

À minha amada esposa que soube compreender minha ausência e me incentivar em todo tempo;

À minha linda filha que, desde a sua concepção até os seus 6 meses atuais, tem sido uma renovação para minhas forças.

À minha mãe, irmãos e a toda família que, com muito carinho,apoio e compreensão, não mediram esforços para que chegasse ao fim desta grande jornada. Amo vocês...

Ao meu orientador, Cláudio de Souza Baptista, pelo incentivo desde a graduação, sendo o primeiro professor a me estimular para a pesquisa acadêmcia. Sou grato por todo o aprendizado transmitido, seja por palavras ou atitudes, durante essa etapa em minha vida. Não posso esquecer dos "puxões de orelhas" que foram necessários para alcançar a qualidade de uma pesquisa de nível acadêmico. Minha gratidão...

Ao Laboratório de Sistemas de Informação, onde pude crescer como profissional e pesquisador. Em especial, aos colegas Anderson Almeida, Maxwell Guimarães e Daniel Leite que contribuíram diretamente com esta pesquisa, seja nas publicações realizadas ou na trocas de experiências...

Não poderia deixar de tecer meus sinceros agradecimentos à Universidade Estadual da Paraíba que, como instituição formadora do saber, mantém uma política de incentivo e capacitação aos seus servidores...

Aos professores e funcionários do Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande por contribuir na minha formação...

Aos professores Cláudio Campelo e Fábio Andrade por compor à banca examinadora e pelas valorosas contribuições a este trabalho.

Conteúdo

Lista de Símbolos

NLP - *Natural language processing*

NER - *Named Entity Recognition*

JSON - *JavaScript Object Notation*

POS Tagging ou POST - *Part-of-Speech Tagging*

KNN - *K-Nearest Neighbourhood*

CRF - *Condition Random Field*

API - *Application Programming Interface*

HTML - *HyperText Markup Language*

SMO - *Sequential Minimal Optimization*

GeoSEn - *Geographic Search Engine*

IBGE - *Instituto Brasileiro de Geografia e Estatística*

GIR - *Geographic Information Retrieval*

IR - *Information Retrieval*

POI - *Point Of Interest*

URL - *Uniform Resource Locator*

SGBD - *Sistema de Gerenciamento de Banco de Dados*

SIG ou GIS - *Geographic Information System*

SLD - *Styled Layer Descriptor*

Lista de Figuras

Lista de Tabelas

Lista de Códigos Fonte

Capítulo 1

Introdução

Com a rápida evolução da Web ocorrida nos últimos anos, o conteúdo digital não estruturado também cresceu drasticamente devido à quantidade de usuários e à forma interativa com a qual estão utilizando a Internet. Somente no Brasil, já são mais de 94,2 milhões[1] de pessoas que usam a Internet. Considerando o acesso à Internet no mundo, estatísticas apontam mais de 2,4 bilhões[2] de pessoas. O fato é que estamos vivenciando uma era de conectividade social, onde as pessoas estão ficando cada vez mais entusiasmadas com a forma de interagir, compartilhar e colaborar através de redes sociais, comunidades on-line, blogs, wikis e outras mídias colaborativas on-line.

A crescente interatividade entre os serviços oferecidos na Web e os seus usuários gera uma enorme quantidade de informação. Com esta nova forma de usar a Web, chamada de Web 2.0, os usuários não navegam simplesmente na Web, eles contribuem ativamente com o seu conteúdo através das aplicações [1], colaborando assim para a formação de uma inteligência coletiva [2]. A Web 2.0 proporcionou uma proliferação de informação não estruturada através de blogs, fóruns de discussões, sites de avaliação de produtos on-line, microblogs e redes sociais das mais diversas, trazendo assim novos desafios e oportunidades na busca e na recuperação da informação [3]. Essa inteligência coletiva se espalhou para diversas áreas, especialmente nas relacionadas com a vida cotidiana, tais como comércio, turismo, educação

[1]Pesquisa realizada pelo IBOPE Media, realizada no terceiro trimestre de 2012. Fonte: http://www.ibope.com.br/

[2]Dado referente ao ano de 2012 divulgado pelo Internet World Stats. Disponível em: http://www.internetworldstats.com/stats.htm

e saúde, fazendo com que a Web Social expanda exponencialmente [4]. Deste modo, compreender o que as pessoas estão pensando ou suas opiniões é fundamental para as tomadas de decisões, principalmente neste contexto em que as pessoas expressam seus comentários de forma voluntária no intuito de cooperar umas com as outras.

No contexto do comércio eletrônico, muitas pessoas utilizam a Internet para verificar opiniões ou avaliações de outras pessoas antes de comprar um produto. A procura por comentários e avaliações de produtos em sites é uma prática muito comum, uma vez que essas opiniões, sejam positivas ou negativas, influenciam muito na decisão do comprador. Neste sentido, do ponto de vista do comprador, os comentários positivos lhe deixam mais confortável na decisão da compra; por outro lado, do ponto de vista do vendedor ou fabricante, os comentários negativos o direcionam no melhoramento do seu produto ou serviço [5]. O fato é que as opiniões têm um fator decisivo na hora da compra. No entanto, a quantidade de comentários tem crescido rapidamente, tornando impossível a análise manual de todos os dados. Em alguns produtos mais populares, o número de comentários/avaliações chega a centenas ou mesmo milhares. Tanto os clientes quanto os fabricantes e vendedores correm o risco de deduzir conclusões equivocadas mediante leituras reduzidas de comentários, proporcionando assim alguma decisão tendenciosa [6]. Desse modo, a automatização da descoberta de opinião e sumarização é extremamente necessária.

Compreender os sentimentos das informações produzidas na Web de forma automática não é uma tarefa trivial, tratando-se de um problema de Processamento de Linguagem Natural [7]. As informações normalmente não estão estruturadas, pois foram produzidas por humanos para a interpretação humana. Então, para extrair o sentimento de um texto é necessário compreender a maioria das regras explícitas e implícitas, regulares e irregulares, sintáticas e semânticas da própria língua/idioma.

Segundo Liu [8], a análise de sentimentos[3], também conhecida na literatura por mineração de opinião, é o campo de estudo que analisa as opiniões das pessoas, sentimentos, avaliações, atitudes, e emoções a favor das entidades, tais como produtos, serviços, organizações, indivíduos, questões, eventos, tópicos e seus atributos. Liu [7] destaca que é possível categorizar informações de texto como fatos ou opiniões [7]. Um fato pode ser dito como

[3]Na literatura científica a análise de sentimentos é conhecida também por outros termos, tais como extração de opinião, mineração de sentimento, análise de subjetividade, análise da emoção e mineração da revisão

uma informação com caráter objetivo sobre alguma entidade, algum evento, algum dado ou alguma de suas propriedades. Já a opinião apresenta um sentido subjetivo expresso por algum indivíduo ou grupo. Na maioria das aplicações, é preciso analisar as opiniões de um grande número de pessoas. Desta forma, a sumarização das opiniões é desejada. Uma forma comum de realizar a sumarização é através da classificação da opinião do objeto em positiva, negativa ou neutra. Este tipo de classificação é conhecido na literatura por classificação da polaridade do sentimento ou classificação da polaridade [9]. Assim, para obter a sumarização de um conjunto de opiniões acerca de um objeto basta quantificar as polaridades das opiniões analisadas, obtendo, desta forma, uma orientação sobre o sentimento geral daquele objeto avaliado.

Sabe-se que já há diversas aplicações que realizam análise de sentimentos, contudo ainda não há uma solução de análise de sentimentos que considere o fator espaço-temporal, isto é, considerar a localização geográfica da fonte da informação ou da informação e possíveis mudanças de opinião ao longo do tempo.

O fator temporal é utilizado para acompanhar as mudanças de opinião, pois, em algumas aplicações, a detecção de mudança de opinião pode auxiliar nas tomadas de decisões em tempo hábil, como no caso das campanhas político-partidárias, cujas alterações de opinião possibilitam mudança de estratégias de marketing. Outra aplicação pode ser em empresas que desejam saber se uma determinada campanha publicitária realizada obteve os resultados esperados.

No caso do fator espacial, a utilização de mineração de opinião georreferenciada pode ser muito importante em aplicações que necessitem compreender o sentimento das opiniões segmentadas de acordo com a localização geográfica da opinião. Por exemplo, um partido político poderia adaptar suas campanhas eleitorais focando em determinadas regiões cuja análise constatou fragilidade. Portanto, este tipo de análise de sentimento espacial possibilita que o agente tomador de decisão considere as demandas regionais.

O objetivo deste trabalho é propor uma abordagem de análise de sentimentos que considera os fatores temporal e espacial em microtextos. Para demonstrar como esses dois fatores podem ser utilizados em uma análise de sentimentos, foram coletadas cerca de 300 mil mensagens do Twitter[4] (tweets) relacionadas com a Copa das Confederações de 2013 no Brasil.

[4] www.twitter.com

Para a realização de análise de sentimentos nos textos, foram implementados e avaliados três algoritmos de análise de sentimentos. O algoritmo que obteve a melhor índice no reconhecimento da polaridade foi utilizado na ferramenta proposta neste trabalho. No âmbito do fator espacial, foram utilizados técnicas de Recuperação de Informação Geográfica (Geographic Information Retrieval - GIR) para detecção de referências geográficas nos tweets através das evidências textuais.

Assim, mediante a detecção das polaridades dos sentimentos e a inferência geográfica realizada nos tweets, foi possível gerar mapas de calor que possibilitem uma análise mais precisa, espacial e temporal, quanto ao sentimento expresso pela população em relação ao tema coletado. Técnicas de Extração e Recuperação da Informação são utilizadas para sumarizar os sentimentos detectadas em microtextos, oferecendo suporte ao tomador de decisão.

As principais contribuições desta pesquisa são:

- Implementação e comparação de técnicas de classificação de sentimentos, que quando aplicadas em microtextos escritos no idioma português, apresentaram resultados de acurácia, precisão e revocação considerados satisfatórios;
- Proposta de sumarização dos sentimentos explorando a dimensão espacial e temporal;
- Utilização de mapas de calor para sumarizar o sentimento através da visualização espacial nas regiões geográficas. Esta é uma ma alternativa de sumarização que ainda não foi explorada na literatura;

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo geral desta pesquisa é desenvolver uma técnica de análise de sentimentos considerando as dimensões espacial e temporal.

1.1.2 Objetivos Específicos

Os objetivos específicos são:

- implementar e avaliar técnicas de mineração de opinião aplicadas em microtextos escritos em português;

- utilizar técnicas de Recuperação de Informação Geográfica para identificar as regiões geográficas nos microtextos;

- possibilitar a sumarização da polaridade do sentimento detectado, explorando as dimensões espacial e temporal;

1.2 Relevância

Desde o início do ano de 2000, pesquisas envolvendo análise de sentimentos tem sido a área de pesquisa mais ativa no campo de Processamento de Linguagem Natural (Natural Language Processing- NLP) [10]. Além disso, a análise de sentimentos está sendo também amplamente estudada em mineração de dados, mineração da Web e mineração de texto [8]. O interesse nesta área de pesquisa deve-se ao crescimento dos meios de comunicação social na Web, como os comentários/revisões, discussões em fóruns, blogs, microblogs e redes sociais, como o Twitter e Facebook, que proporcionou um volume de dados opinativos armazenados em formato digital nunca visto na história da humanidade. Assim, a análise de sentimentos realizada sobre estes dados constitui uma fonte importante e rica para se entender e se antecipar às expectativas e frustrações das pessoas a respeito de um produto, um serviço ou mesmo sobre pessoas e fatos. É importante observar que as opiniões sobre os mais diversos temas expressos pelos usuários da Web são feitas de forma espontânea, gratuita e em tempo real.

Haja vista a oportunidade de capturar as opiniões do público em geral sobre algum tema, a análise de sentimentos tem despertado o interesse crescente, tanto no seio da comunidade científica, ainda com muitos desafios abertos, como também no mundo dos negócios, devido aos benefícios de compreender o sentimento das pessoas de forma automática para as tomadas de decisões. A análise de sentimentos tem sido usada para diversas aplicações e diversos propósitos:

- análise de empresas na bolsa de valores [11; 12], cujo objetivo é identificar o humor do mercado em relação às empresas negociadas na bolsa de valores baseado nas opiniões

dos analistas (jornais e bloggers), com o intuito de identificar a tendência dos preços das mesmas;

- análise de produtos [3; 6], na qual uma empresa ou mesmo usuários têm interesse na opinião dos consumidores sobre um determinado produto. Esta análise pode ser feita de uma forma geral sobre os comentários ou através da extração e sumarização das características do produto;

- análise de lugares [13], onde um turista pretende viajar pode utilizar as opiniões de terceiros para planejar o roteiro da viagem, evitando assim passeios desinteressantes;

- análise de políticos [14; 15] ou assuntos de política [16], em que os eleitores podem identificar qual a opinião de outros eleitores sobre um determinado candidato;

- análise de filmes [17] e jogos, onde também é possível realizar a mineração de opinião [6];

Em Bjørkelund et al. [13], a utilização de mapas dá-se apenas para auxiliar os usuários na detecção da região de seus interesses. Não há possibilidade de realizar operações espaciais para detectar, por exemplo, quais os hotéis ou regiões que sofreram mudanças de opinião em um determinado período. Já em Dias [18], a informação temporal não é considerada na análise, sendo inclusive uma das sugestões de trabalhos futuros do autor. A análise de opinião temporal é relativamente um campo de pesquisa recente, sendo conhecida por outras nomenclaturas como mineração de opinião temporal, análise de opinião time-aware, mineração de mudança de opinião e rastreamento de opinião. Trata-se de um processo de monitoramento e possível detecção de mudanças de opiniões sobre um determinado tema em um período de tempo específico, e pode ser visto como uma extensão à mineração de opinião [19]. Desta forma, esta pesquisa trata de temas recentes no campo de análise de sentimentos ao explorar as dimensões espacial e temporal em microtextos, contribuindo para a compreensão do sentimento detectado na Web de forma automática através dos mecanismos de sumarização da opinião propostos neste trabalho.

1.3 Trabalhos Publicados

As seguintes publicações resultaram diretamente desta pesquisa:

- Temporal Analysis of Sentiment in Tweets: a Case Study with FIFA Confederations Cup in Brazil.

  Conferência: DEXA 2014 - 25th International Conference on Database and Expert Systems Applications

  Local: Munich, Germany

- A comparison of SVM versus Naive-Bayes Techniques for Sentiment Analysis in Tweets: a Case Study with the 2013 FIFA Confederations Cup

  WebMedia 2014 - XX Simpósio Brasileiro de Sistemas Multimídia e Web

  Local: João Pessoa, Paraíba

1.4 Organização Estrutural

O restante desta dissertação está organizado da seguinte forma: no Capítulo 2, é apresentada a fundamentação teórica desta pesquisa, que aborda os temas Análise de Sentimentos, Recuperação de Informações Geográficas e Métricas de Avaliação em Sistema de Recuperação da Informação. No Capítulo 3, são mostrados os trabalhos relacionados a esta pesquisa. A abordagem de análise de sentimentos espaço-temporal em microtexto é apresentada no Capítulo 4. No Capítulo 5, são descritos os experimentos realizados através de um estudo de caso e as validações das técnicas de análise de sentimentos e georeferenciamento. Finalmente, no Capítulo 6, estão descritas as conclusões e proposições para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

O objetivo deste capítulo é abordar os principais conceitos e tecnologias utilizadas nesta pesquisa. Para tanto, ele está subdividido em duas grandes seções, que tratam, respectivamente, sobre Análise de Sentimentos e Recuperação de Informações Geográficas. A última seção trata das métricas de avaliação utilizadas em sistemas de RI.

2.1 Análise de Sentimentos

Liu [8] define a análise de sentimento como o campo de estudo que analisa a opinião das pessoas, sentimentos, avaliações, atitudes e emoções em entidades como produtos, serviços, organizações, indivíduos, questões, eventos, tópicos e seus atributos. Na literatura, outros termos são encontrados para referir-se a análise de sentimentos, como mineração de opinião, extração de opinião, mineração de sentimento, análise de subjetividade, análise da emoção, mineração da revisão [8; 9; 20].

Devido a possibilidade de capturar as opiniões das pessoas sobre algum tema de forma automática, a análise de sentimento têm despertado bastante interesse, tanto da comunidade científica, devido aos problemas ainda em aberto, quanto das empresas, devido aos benefícios em compreender os sentimentos das pessoas em tempo real e de forma automática, para auxiliar nas tomadas de decisões.

Comercialmente, a mineração de opinião têm se tornado cada vez mais popular no ambiente empresarial impactando os negócios das através das aplicações conhecidas por Business Intelligence(BI). Business Intelligence pode ser definido como um conjunto de técnicas

computacionais utilizadas para extrair inteligência a partir de dados sobre um determinado negócio. Um dos objetivos incluem a compreensão e análise dos pontos positivos e negativos das empresas e sua relação com os clientes. Analisar parâmetros objetivos, como peso de um produto, tamanho ou custo é uma tarefa bem mais simples de ser realizada. No entanto, analisar aspectos subjetivos dos produtos como o design do produto ou a sua usabilidade envolve compreender as opiniões dos clientes. A análise de sentimentos pode ajudar a responder a esses tipos de preferências subjetivas dos clientes [19].

A análise de sentimentos é uma área de pesquisa recente, que utiliza técnicas avançadas de mineração de texto, aprendizagem de máquina, recuperação de informação e Processamento de Linguagem Natural (NLP) para processar grandes quantidades de conteúdos não estruturados gerados por usuários, principalmente nas mídias sociais [21]. Dessa forma, o objetivo da análise de sentimentos é extrair a opinião e o conhecimento subjetivo de textos on-line, formalizar esse conhecimento descoberto e analisá-lo para uso específico [7].

2.1.1 Definição de Opinião

Uma opinião é composta por pelo menos dois elementos chave: um alvo e e um sentimento s sobre a entidade e. Algebricamente, uma opinião é definida como uma tupla [8]

$$O = (e, s),$$

onde o alvo e pode ser qualquer entidade ou aspecto/característica da entidade, como por exemplo uma pessoa, evento ou produto que deseja-se expressar o sentimento s. Já o sentimento s representa a opinião expressa sobre e, podendo ser um sentimento positivo, negativo ou neutro, ou ainda um ponto de escala que expressa a intensidade do sentimento sobre a entidade, como por exemplo uma escala de 1 a 5, onde 1 representa um sentimento muito negativo e 5 representa um sentimento muito positivo.

Considerando que:

- a opinião é expressa por emissor de opinião h, que representa a fonte de opinião ou o detentor do sentimento;
- a opinião é realizada em um instante de tempo t;

- a entidade e pode ser considerada por diferentes propriedades/perspectivas, características (feature), ou simplesmente aspectos a;

uma definição formal de opinião mais completa pode ser dada pela quíntupla [8]:

$$O = (e_i, a_{ij}, s_{ijkl}, h_k, t_l), \tag{2.1}$$

onde:

- e_i representa o nome de uma entidade;
- a_{ij} é um aspecto/característica da entidade e_i, sendo um elemento opcional utilizado apenas quando deseja-se um nível de detalhamento maior das entidades;
- s_{ijkl} é a polaridade do sentimento sobre o aspecto a_{ij} da entidade e_i;
- h_k é um emissor da opinião; e
- t_l é um instante que a opinião foi expressa por h_k.

Em muitas aplicações, a utilização do conceito de aspectos é necessária para detalhar os sentimentos sobre os diversos aspectos avaliados. Por exemplo, no âmbito do e-commerce os clientes necessitam conhecer opinião de outros clientes não apenas sobre o produto em geral, mas das características do produto avaliado [6]. Considere a seguinte avaliação de uma câmera digital realizada em um site de e-commerce pelo usuário Y no dia 09 de junho de 2014: "O Smartphone X apresenta uma câmera com excelente qualidade de imagem, além de um design sofisticado. No entanto, o visor é bastante pequeno". Neste caso, informar se o sentimento sobre a câmera é positivo ou negativo, depende muito do ponto de vista ou simplesmente do aspecto considerado. Assim, o sentimento pode ser analisado conforme o aspecto considerado, ou seja,

- (Smartphone X, câmera, positivo, Y, 09/06/2014)
- (Smartphone X, design, positivo, Y, 09/06/2014)
- (Smartphone X, visor, negativo, Y, 09/06/2014)

Ao considerar os aspectos na análise de sentimento deve-se observar também os aspectos implícitos, como por exemplo "A câmera é muito cara", cujo aspecto avaliado é o "preço" da câmera. Outro fator a ser considerado são os adjetivos utilizados para descrever os aspectos dos produtos, como por exemplo "A câmera tem uma ótima resolução" onde o aspecto avaliado é a imagem da câmera. A atividade de detectar automaticamente os aspectos de uma entidade e é também objeto de pesquisa [6; 8].

A polaridade de um sentimento, ou simplesmente a orientação do sentimento, refere-se à classificação do sentimento em positivo, negativo ou neutro. Muitos trabalhos de classificação de sentimento preferem a classificação binária, que consideram apenas os sentimentos positivos e negativos dos documentos opinativos avaliados [9]; no entanto, a polaridade de um sentimento ou polaridade da classificação não necessariamente está associada a classificação binária. Na literatura, o termo polaridade do sentimento pode também estar associado à categorização multi-classe que descreve o grau de positividade ou negatividade do sentimento avaliado (intensidade do sentimento). A classificação "neutra" de um sentimento é utilizada para indicar que não há um sentimento expresso no documento/texto avaliado, tratando assim de um texto objetivo/informativo.

2.1.2 Tarefa de Análise de Sentimentos

Dado um conjunto de documentos D, têm-se que o objetivo da análise de sentimento é descobrir todas as quíntuplas $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ em um dado documento $d \in D$. Desta forma, a principal atividade da classificação do sentimento é obter a opinião predominante sobre as expressões textuais do documento d [7].

O processo de análise de sentimento pode ser definido em três grandes tarefas [8; 20], conforme ilustra a Figura 2.1.

A tarefa de Identificação pode incluir além do reconhecimento das entidades e seus aspectos (quando necessário), também o reconhecimento de sentenças subjetivas/opinativas. A identificação de entidades e seus aspectos é um problema de Reconhecimento de Entidades Nomeadas (NER) [22; 23]. Reconhecer as entidades e seus aspectos por si só representa um grande problema a ser resolvido, dependendo do documento a ser considerado e de seu nível de estruturação, pois pessoas podem escrever sobre as mesmas entidades, mas de formas diferentes. Um outro problema no reconhecimento de entidades, está relacionado com as

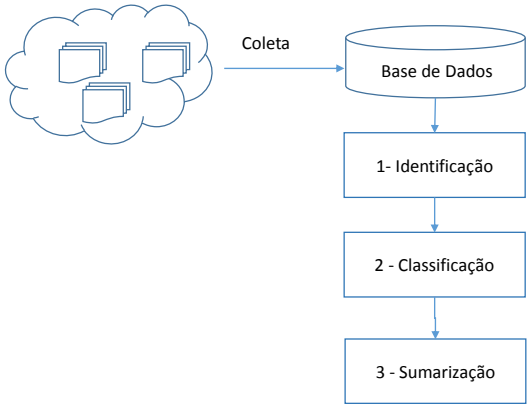


Figura 2.1: Processo de análise de sentimentos.

co-referências e utilização de pronomes relacionados com as entidades. Desta forma, uma correta análise de sentimentos precede de um correto processamento da linguagem natural (NLP). O reconhecimento de entidades em documentos do tipo jornais, blogs e posts requer um maior esforço, uma vez que as entidades avaliadas são desconhecidas. Algumas aplicações de análise de sentimentos utilizam-se de documentos cujas entidades avaliadas já são conhecidas/pré-definidas, como por exemplo, a utilização de documentos que contêm as avaliações de produtos; e neste caso o esforço está exclusivamente na identificação dos aspectos (características dos produtos) [6].

Outra tarefa relacionada com a identificação e que pode ser realizada antes da atividade de classificação da polaridade do sentimento, visando inclusive uma melhor performance na análise do sentimento, é o discernimento entre um conteúdo objetivo (descreve fatos) e subjetivo (contém opinião) [24].

A segunda tarefa da análise de sentimento está relacionada à classificação da polaridade do sentimento, ou simplesmente classificação do sentimento. A classificação da polaridade é a principal atividade em aplicações de análise de sentimentos, representando vários desafios a serem tratados, tais como:

- identificação de ironias/sarcasmos, uma vez que estas invertem o sentido do sentimento exposto;
- as opiniões podem ser explícitas ou implícitas (através de comparações ou citações indiretas);

- subjetividade das opiniões: pontos de vistas diferentes quanto as opiniões. Seres humanos divergem quanto a polaridade do sentimento (dificilmente o consenso é maior que 75%) [9];

- adjetivos ou termos que expressam uma polaridade em um determinado contexto pode representar a polaridade oposta em outro contexto;

Para classificação da polaridade várias abordagens já foram propostas. Na seção 2.1.4, as principais técnicas e que foram utilizadas neste trabalho para identificação da polaridade de sentimentos são abordadas;

Na maioria das aplicações, que utilizam-se da análise de sentimento, estão interessadas na opinião de várias pessoas, pois analisar a opinião de apenas uma pessoa não é suficiente para a tomada de decisões [8]. Neste caso, as aplicações de análise de sentimentos utilizam-se da terceira atividade, que é a de sumarização da opinião. Esta etapa é responsável pela criação de métricas e sumários que representam o sentimento geral ou de um grupo de pessoas sobre uma determinada entidade ou aspectos aspectos da entidade. Para aplicações que utilizam-se de aspectos das entidades avaliadas, como revisões de produtos, é comum a exibição do sentimento detectado em cada aspecto, seja de forma gráfica (exemplos Google Shopping ou Bing), como ilustra a Figura 2.2, ou de forma quantitativa, como ilustra a Figura 2.3. Em ambos os casos, através da detecção dos termos mais citados, exibi-se também as principais opiniões de acordo com um ranking das opiniões mais citadas [6].

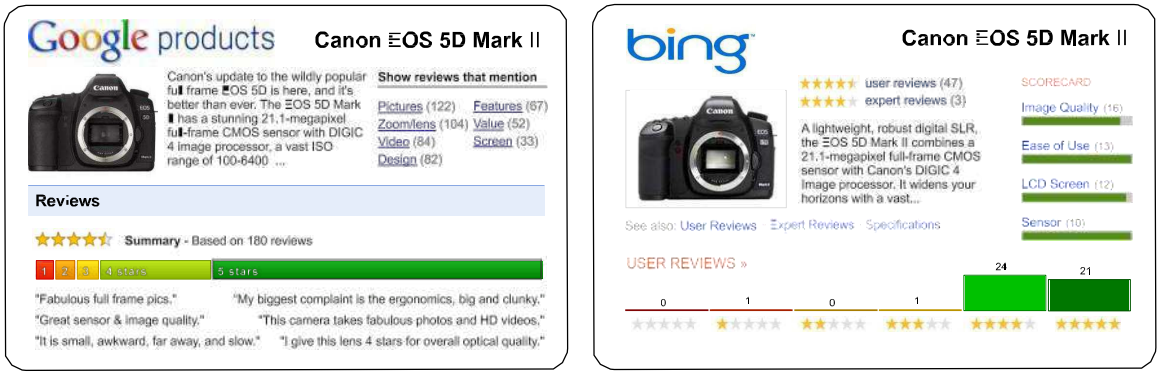


Figura 2.2: Exemplo de sumarização de opinião realizada pelas ferramentas de busca do Google e Bing, respectivamente (Figura retirada de [20]).

*Digital_camera_*1:
Feature: **picture quality**
Positive: 253
<individual review sentences>
Negative: 6
<individual review sentences>
Feature: **size**
Positive: 134
<individual review sentences>
Negative: 10
<individual review sentences>
...

Figure 1: An example summary

Figura 2.3: Exemplo de sumarização de opinião baseada em aspectos (Figura retirada de [6]).

Outra forma de sumarizar a opinião é através de uma análise temporal que associa os sentimentos positivos e negativos ao longo do tempo. Quando existe uma localização geográfica da opinão detectada é possível ainda apresentar o sentimento de forma geográfica, como ilustra as Figuras 2.4 e 2.5.

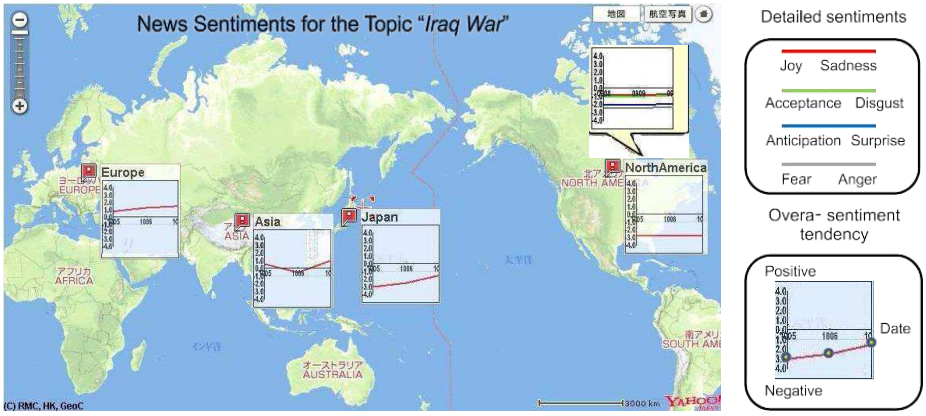


Figura 2.4: Exemplo de sumarização do sentimento geográfico (Figura retirada de [25]).

Assim, a tarefa de sumarização da opinião, objetiva identificar a opinião média ou prevalecente detectada [20], facilitando a tomada de decisão.

2.1.3 Arquitetura de um Sistema de Análise de Sentimentos

A Figura 2.6 ilustra uma arquitetura genérica de um sistema de análise de sentimentos.

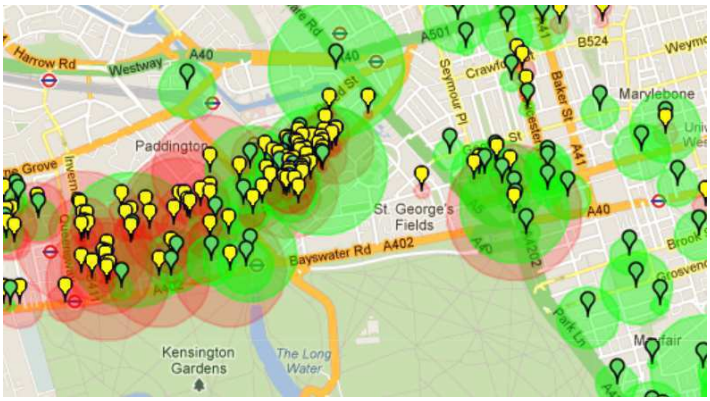


Figura 2.5: Exemplo de sumarização do sentimento geográfico com cluster. Figura retirada de [13]

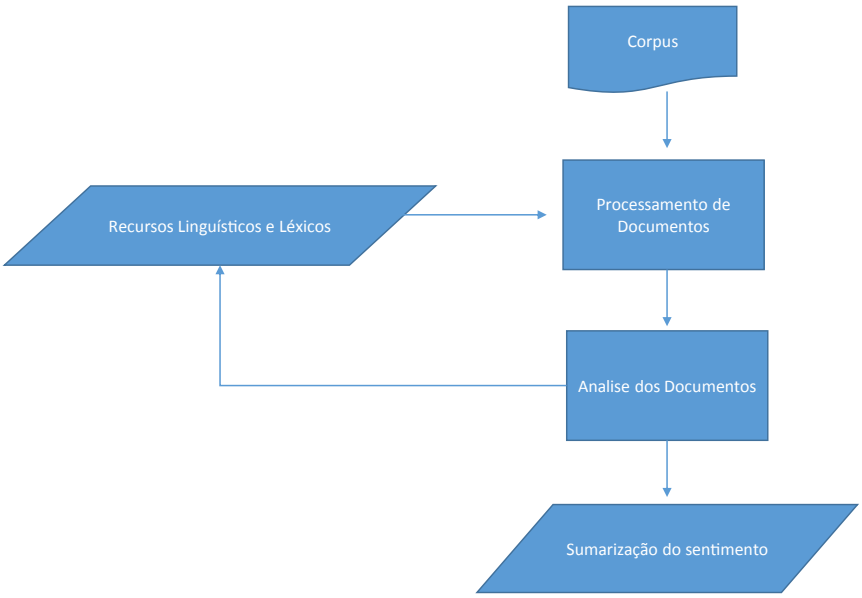


Figura 2.6: Exemplo de sumarização do sentimento geográfico com cluster (Adaptado de [10]).

A entrada do sistema é o conjunto de textos contidos nos documentos em algum formato digital (PDF, HTML, XML, Doc, etc). Assim como os sistemas de Processamento de Linguagem Natural (NLP), os textos passam na fase de pré-processamento utilizando uma variedade de arcabouços linguísticos como:

Stemming: método para redução de um termo ao seu radical, removendo as desinências, afixos, e vogais temáticas;

Part Of Speech Tagging: identificação das classes gramaticais das palavras do texto;

Filtragem: Remoção de stopwords e de termos que não interferem na identificação da polaridade. Em tweets, por exemplo, pode-se remover URL's, nomes de usuários do Twitter (inicia com @) e palavras especiais do Twitter (RT, via);

Tokenização dividir o texto em uma lista de termos que o compõe;

A depender do idioma do texto, o sistema poderá utilizar dos recursos linguísticos e léxicos do idioma para facilitar a identificação dos elementos textuais. O principal componente e objeto de estudo de várias pesquisas da área de análise de sentimentos é o módulo de Análise, cujo objetivo é a identificação dos sentimentos contidos nos textos. Por fim, o componente de sumarização é o componente que agrega valor às ferramentas de análise de sentimentos, sendo este módulo responsável por prover mecanismos de compreensão do sentimento geral detectado nos documentos.

2.1.4 Abordagens para detecção da polaridade do sentimento

Basicamente, as pesquisas envolvendo análise de sentimentos considera três níveis de granularidade:

- Nível de Documentos: considera que o documento inteiro trata sobre a opinião de uma única entidade. Desta forma o documento expressa um sentimento positivo ou negativo sobre a entidade considerada. Esta tem sido a forma mais amplamente analisada na literatura [10; 9; 6; 26; 13; 8].

- Nível de Sentenças: Neste nível de análise, a classificação do sentimento é realizada por sentenças, nas quais o documento é dividido [27]. Neste caso, considera-se que há

várias entidades a serem avaliadas e assume-se que há sentenças que são opinativas e sentenças não opinativas.

- Nível de Aspectos: A análise em nível de aspecto permite detalhar o alvo do sentimento, de tal forma que possam ser detectados seus pontos fortes e fracos, analisando os aspectos da entidade avaliada. Assim, a análise da entidade é obtida através da análise dos aspectos/características da entidade.

Quanto às abordagens utilizadas para detecção do sentimento, as principais estão divididas em quatro técnicas: aprendizagem de máquina (classificação ou regressão); análise léxica (análise linguísticas ou baseadas em dicionários), que utilizam dicionários de palavras com sentimentos já identificados; estatísticas que avaliam a co-ocorrência de termos; e semânticas, que definem a polaridade de palavras em função de sua proximidade semântica com outras de polaridade conhecidas. As diferentes técnicas podem ser combinadas para obter melhores resultados.

A seguir serão descritas as duas primeiras abordagens que, além de serem predominantes nas aplicações da literatura [20], também foram utilizadas neste trabalho. Para maiores detalhes sobre as abordagens consultar [8; 9; 20; 28].

Abordagem Léxica

A abordagem baseada em dicionário é também denominada de léxica ou linguística. O aspecto central desta abordagem é o uso de léxicos (dicionários) de sentimentos, que são compilações de palavras ou expressões de sentimento associadas à respectiva polaridade [8; 29]. Um dos métodos mais utilizados na abordagem linguística é o da co-ocorrência entre alvo e sentimento, que não leva em consideração nem a ordem dos termos dentro de um documento (bag of words), nem suas relações léxico-sintáticas.

Para a classificação do sentimento em um texto, basta que exista uma palavra de sentimento, onde sua polaridade é dada por um léxico de sentimentos. Esse método é bastante utilizado quando as sentenças avaliadas contêm apenas uma entidade a ser avaliada. O método por co-ocorrência apresenta bons resultados quando o nível de análise textual é de granularidade pequena, pois a palavra detentora do sentimento está próxima à entidade.

Abordagem baseada em Aprendizado de máquina

Aprendizado de Máquina é uma área de Inteligência Artificial cujo objetivo é a construção de sistemas capazes de adquirir conhecimento de forma automática. O conhecimento é adquirido através de experiências acumuladas baseado da solução bem sucedida de problemas anteriores. Dentre as técnicas de aprendizado, destacam-se as técnicas de aprendizado supervisionado, as quais utilizam-se de dados rotulados previamente. No aprendizado supervisionado, os modelos de classificação são bastantes utilizados em problemas de classificação da polaridade do sentimento, especialmente os classificadores de Naive-Bayes e o SVM, abordados neste trabalho.

Naive-Bayes: O classificador Naive-Bayes utiliza técnicas que trabalham com a modelagem da incerteza através de probabilidades, considerando as entradas independentes entre si. No entanto, mesmo para classes de problemas que tenham atributos altamente dependentes, o Naive-Bayes apresenta ótimos resultados [9; 30; 31]. Isso ocorre devido ao fato de que a independência condicional dos atributos não é uma condição necessária para a otimalidade do Naive-Bayes [32].

O algoritmo Naive-Bayes basicamente traz consigo o mesmo fundamento matemático do Teorema de Bayes [33]. Aplicando esse teorema ao contexto de classificadores de textos, pode-se calcular a probabilidade de um texto pertencer a uma categoria, conforme descrito na equação 2.2:

$$P(c|t) \quad = \quad \frac{P(c)P(t|c)}{P(t)} \tag{2.2}$$

onde:

$P(c)$ = Probabilidade da categoria ocorrer.

$P(t)$ = Probabilidade do texto ocorrer.

$P(t|c)$ = Probabilidade do texto ocorrer, dado que a categoria ocorreu.

$P(c|t)$ = Probabilidade do texto pertencer a categoria, dado que ele ocorreu.

O termo $P(t|c)$ é calculado levando em consideração a probabilidade condicional de cada palavra que compõe o tweet ocorrer, dado que a categoria ocorreu. Esse termo poderia ser escrito da seguinte forma:

$$P(t|c) \quad = \quad \prod_{1 \leq k \leq n} P(t_k|c) \tag{2.3}$$

onde:

$P(t_k|c)$ = = Probabilidade do termo k ocorrer, dado que a categoria ocorreu.

n = Quantidade de termos do tweet.

Suporte Vector Machine: SVM [34] é um método de aprendizagem de máquina com base na minimização do risco estrutural envolvido na criação do hiperplano de alta dimensão para a separação de classes, sendo altamente eficaz para categorização de textos [21]. SVM procura um hiper-plano representado por vetores que divide os vetores de treinamento nas classes desejadas (ex. positiva e negativa). A ideia básica por trás do algoritmo de treinamento é encontrar um hiper-plano, representado pelo vetor $\overrightarrow{w}$ que separa os vetores de documentos em classes por um espaço claro que seja tão amplo quanto possível. Isso faz com que $c_j \in C$ categorize corretamente o documento $\overrightarrow{d}_j$. O problema de encontrar este hiperplano pode ser traduzido em um problema de otimização, apresentando a seguinte solução [21]:

$$\overrightarrow{w} = \sum_j \alpha_j c_j \overrightarrow{d}_j, \ \alpha_j \geq 0,$$

onde α_j é obtido através da solução do problema de otimização. Para os casos em que α_j é maior que zero, os $\overrightarrow{d_j}$ são chamados de vetores de suporte, pois eles são os únicos vetores de documento que contribuem para $\overrightarrow{w}$. Com relação às instâncias de testes, a classificação se restringe a determinar em qual lado do hiperplano $\overrightarrow{w}$ elas irão ficar.

2.2 Recuperação de Informação Geográfica

Recuperação de Informação Geográfica (Geographic Information Retrieval - GIR) pode ser vista como um ramo especializado da tradicional Recuperação da Informação (Information Retrieval - RI), incluindo todas suas áreas de pesquisa. A ênfase está na indexação de dados geográficos e espaciais [35] e as principais atividades envolvem a extração e resolução de nomes de locais em textos não estruturados. Uma das atividades de um sistema de GIR é inferir de forma automatizada localizações geográficas associadas em documentos.

Ignorando a existência de metadados contidos am alguns documentos, há várias maneiras de deduzir informações geográficas com base, por exemplo, no conteúdo do documento e links da Web. O processo de detectar localizações geográficas pode ser dividido em duas etapas [36]: extração e mapeamento. Na etapa de extração, são identificados os elementos que fazem referências a localidades geográficas, como por exemplo, nomes de lugares e códigos postais. Na etapa de mapeamento cada referência detectada é associada a uma localidade geográfica válida.

2.2.1 Desafios

A recuperação da informação geográfica requer que nomes de lugares e frases com referências, diretamente ou indiretamente, sejam resolvidos e traduzidos dentro de footprints[1]. No entanto, existem alguns problemas em associar os footprints dos nomes de lugares com identificadores únicos em documentos, pois as referências textuais podem ser ambíguas e imprecisas, induzindo erros nas referências geográficas. Dentre os diversos problemas, pode-se destacar [35]:

- Ambiguidade: diversos lugares podem ter o mesmo nome ou objetos com nomes de lugares;
- Caráter temporal ou convenção cultural: Alguns nomes de lugares mudam de nome o tempo;
- A extensão do território pode ser alterada, como os desmembramentos de regiões ou a unificação de territórios;
- A fronteira da localização pode ser confusa;
- Alguns nomes ou designador espacial denotam uma área ou localização bastante diferente da área atual definida oficialmente;
- Multiplicidade de nomes: O mesmo nome de um lugar pode ser escrito de forma diferente em um texto (alguns de forma oficial e outros usa consenso popular - termo mais conhecido)

[1] footprints são as localizações geográficos que podem ser representadas pelas coordenadas de latitude e longitude, incluindo a extensão geográfica caso necessário.

Segundo Leidner e Lieberman [37], as ambiguidades presentes na linguagem natural e os erros ortográficos nos documentos são as principais dificuldades na detecção de referências geográficas. O tipo de ambiguidade mais relevante é quando muitas entidades não espaciais compartilham os mesmos nomes de entidades espaciais, como por exemplo "Paris"que pode designar uma cidade (Capital da França) ou uma pessoa (Paris Hilton). Ainda de acordo com Leidner e Lieberman [37], detectar referências geográficas não está associado apenas a reconhecer uma localização através de um nome, mas também reconhecer frases geográficas complexas a exemplo de "30 km de Campina Grande".

2.2.2 Detecção de Referências geográficas

O processo de detecção de referências geográficas é composto de forma geral, pelas fases de Pré-processamento, Geoparsing e Geocoding.

Na fase de Pré-processamento as informações textuais são separadas de informações adicionais como metadados, formatação e layouts. Dependendo da origem do documento esta etapa pode abranger desde a simples coleta da informação textual até o tratamento de ruídos textuais, como correção automática da ortografia para facilitar a identificação dos elementos geográficos.

O componente do geoparsing compreende a identificação das referências (diretas e indiretas) geográficas, como os nomes de lugares ou outros termos relacionados com o espaço geográfico. As principais técnicas utilizadas no geoparsing baseiam-se em heurísticas (Gazetter Lookup Based) ou em técnicas de Inteligência Artificial (Rule Based, Reconhecimento de Entidades Nomeadas e Aprendizado de Máquina) [37].

A etapa de geocoding, conhecida também como a resolução de topônimos, é responsável por obter a representação geográfica das referências identificadas, buscando desambiguar as localidades recuperadas, ou seja, associar cada localidade a apenas uma única localização geográfica (footprint). Um algoritmo de ranking dos footprints identificados faz-se necessário para melhor qualificar as localidades segundo o grau de associação com o texto, uma vez que um mesmo nome de uma localização pode estar associado com diversas regiões.

Para converter os lugares geográficos em coordenadas geográficas, no processo de geocoding, são utilizados os serviços de um ou mais gazetteers (ou ontologias geográficas), os quais integram técnicas de representação do conhecimento e ajudam a estabelecer corres-

pondências e relações entre os diferentes domínios de entidades espaciais [38]. O Gazeetter provê um esquema de organização hierárquica, possibilitando alguns tipos de inferências (i.e topológicas ou relações hierárquicas). Gazeetter é considerado uma das melhores formas de representar o mundo geográfico [39].

O reconhecimento correto do contexto geográfico de um documento requer que as tarefas de geoparsing e geocoding sejam feitas corretamente. Devido as constantes modificações em nomes de lugares, seja de forma oficial ou informal, é importante a utilização de Gazetters dinâmicos capazes de abranger as especificidades das localidades [40].

A seguir será apresentada uma ferramenta que, dentre outras atividades relacionadas a GIR, realiza o processo de detectar referências geográficas em documentos textuais.

GeoSEn: GEOgraphic Search ENgine

O GeoSEn é um motor de busca com enfoque geográfico que realiza a indexação geográfica de documentos extraídos da Web, possibilitando a detecção de referências geográficas dos documentos baseado em um conjunto de heurísticas, que atribui um índice de confiança nas localizações detectadas. Este índice de confiança além de ser utilizado para rejeitar a referência caso apresente um índice inferior a um limiar, ele é utilizado também no processo de desambiguação. A Figura 2.7 ilustra a arquitetura do GeoSen.

O processo de reconhecimento de localizações geográficas é composto principalmente pelos módulos de detecção de lugares (geoparsing) e modelagem do escopo geográfico dos documentos (georeferencing). Utilizando esses dois módulos, é possível inferir localizações geográficas citadas em um texto escrito em língua portuguesa numa hierarquia de divisão política, partindo desde o nível menos preciso (países) até um mais preciso (municípios). O GeoSEn trabalha com o escopo geográfico definido pela hierarquia geográfica $cidade \mapsto$ microrregião $\mapsto$ mesorregião $\mapsto$ estado $\mapsto$ região, a qual é utilizada para mensurar o grau de relevância de cada localidade geográfica detectada em relação ao documento. Além de reconhecer os nomes de lugares, o GeoSen trabalha com outros tipos de referências geográficas como CEP e Códigos de Telefone.

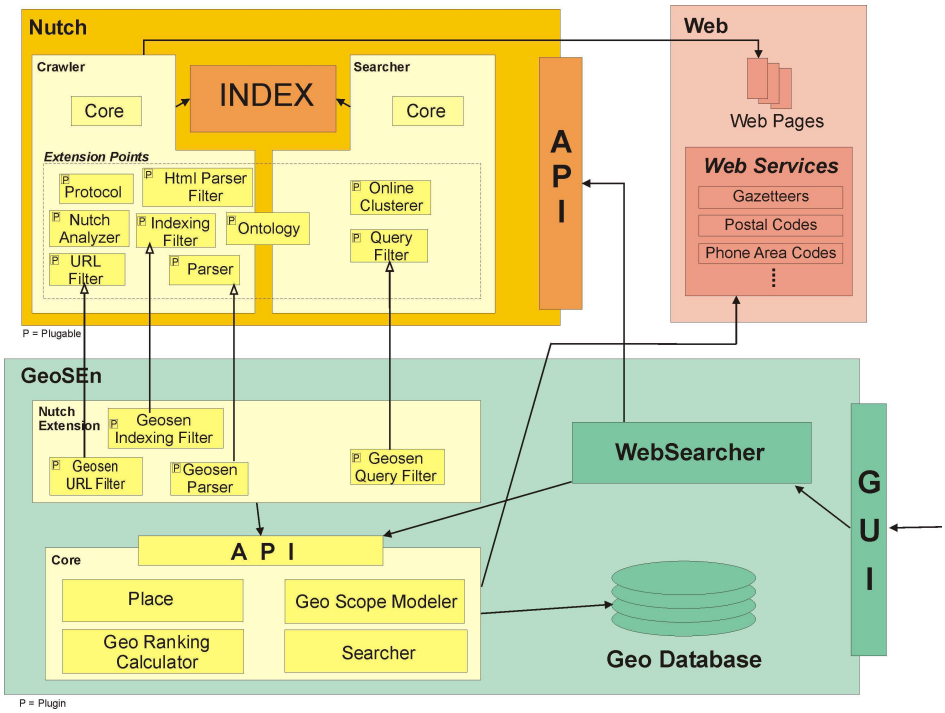


Figura 2.7: Arquitetura do GeoSEn (Imagem extraída [41])

2.3 Avaliação em Sistema de Recuperação de Informação

Na literatura, sistemas de RI, são normalmente avaliados pelas métricas de acurácia, precisão (precision), revocação (recall) e F-Measure [42], definidas, respectivamente, pelas equações 2.4, 2.5, 2.6, e 2.7.

- TP (verdadeiro positivo) indica os verdadeiros positivos, que é definido como o número de pares tweet-sentimento que o sistema identifica corretamente como positivos;
- TN (verdadeiro negativo) vindica os verdadeiros negativos, que é definido como o número de pares tweet-sentimento que o sistema identifica corretamente como negativos;
- FP (falso positivo)indica os falsos positivos, que é definido como como o número de pares tweet-sentimento que o sistema identifica falsamente como positivos; e
- FN (falso negativo) indica os falsos negativos, que é definido como o número de pares tweet-sentimento que o sistema identifica falsamente como negativos.

$$\text{Acurácia} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.4}$$

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{2.5}$$

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{2.6}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recal} \tag{2.7}$$

Especificamente, na área de Análise de Sentimentos, são as métricas utilizadas na literatura para avaliação dos algoritmos de detecção de polaridade dos a sentimentos e serão utilizadas para avaliação dos algoritmos implementados neste trabalho.

A acurácia refere-se a porcentagem de amostras positivas e negativas classificadas corretamente sobre a soma de amostras positivas e negativas. A precisão refere-se a porcentagem de amostras positivas classificadas corretamente sobre o total de amostras classificadas como positivas. Já a revocação (sensibilidade) refere-se a porcentagem de amostras positivas classificadas corretamente sobre o total de amostras positivas. E por fim, a F-measure, também chamada F-score, é uma média ponderada de precisão e revocação.

2.4 Considerações do Capítulo

Este capítulo apresentou a fundamentação teórica deste trabalho. Foram apresentadas os principais conceitos relacionados com análise de sentimentos e aspectos relacionados recuperação de informações geográficas. No próximo capítulo, serão apresentados alguns trabalhos relacionados à pesquisa realizada nesta dissertação.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, serão apresentados os principais trabalhos relacionados a esta pesquisa. A seção 3.1 apresenta os trabalhos que utilizam técnicas de análise de sentimentos. Em seguida, na seção 3.2, são apresentados os trabalhos de análise de sentimentos realizados no idioma português. Na seção 3.2, são apresentados os trabalhos que utilizam sumarização espacial nas abordagens de análise de sentimentos. E, finalmente, na seção 3.4, é apresentado um quadro comparativo contendo as principais características apresentadas nos trabalhos relacionados.

3.1 Abordagens de Análise de Sentimentos

Desde o início do ano de 2000, a análise de sentimentos tem sido uma das áreas mais pesquisadas no campo de Natural Language Processing [8]. A análise de sentimentos tem sido utilizada em diversas aplicações e propósitos: em empresas de bolsa de valores, identificando o humor do mercado baseado nas opiniões de especialistas [11; 12]; em análise de revisões dos consumidores de produtos ou serviços [3; 6]; análise de lugares ou regiões turísticas realizadas através dos comentários dos viajantes [13]; análise de políticos [14; 15] ou assuntos relacionados a política [16].

Descobrir o que as pessoas pensam, segundo Pang e Lee [9], sempre foi objeto de interesse. No contexto da Web Social, através da popularização de plataformas que fornecem acesso a grande quantidade de dados subjetivos, compreender de forma automática a opinião das pessoas sobre algum tema, serviço ou produto tem sido um fator essencial para a tomada

de decisões. Para as organizações, analisar as opiniões das pessoas por meio das mídias sociais significa ampliar as fontes de opinião, tornar mais barata a coleta dos dados e reduzir o tempo de processamento da informação [29].

Atividades relacionadas com a análise de sentimentos envolvem a detecção de conteúdo subjetivo ou opinativo, classificação da polaridade do conteúdo e sumarização dos resultados do sentimento geral das entidades avaliadas. A detecção do sentimento em um texto ocorre em diferentes granularidades: nível de documento, nível de sentença e nível de entidade ou aspectos. Vários métodos já foram propostos para classificar a polaridade do sentimento de um texto, e as principais abordagens utilizadas são baseadas em técnicas de aprendizado de máquina, técnicas de análise semântica, técnicas estatísticas e técnicas baseadas em análise léxica ou dicionário. No entanto, na literatura há uma predominância de utilização das técnicas baseadas em dicionários e as de aprendizado de máquina, sendo que esta última tem se destacado por apresentar melhores resultados [21; 10; 29].

As abordagens em análise de sentimentos que utilizam aprendizagem de máquina implementam algoritmos de classificação como o Naïve-Bayes, Máquina de Vetores de Suporte (SVM), Entropia Máxima, Árvores de Decisões (C4.5), KNN (K-nearest neighbour)e Condition Random Field (CRF) [33]. Uma das principais limitações no uso de aprendizado supervisionado é a necessidade de dados rotulados para treino e testes. No idioma inglês existem vários dados disponíveis já rotulados referentes a comentários de filmes [21], produtos [6] e hotéis [13], e que podem ser utilizados para treinamento e testes nestes domínios específicos. No entanto, outros idiomas e domínios carecem de dados rotulados para treinamento dos modelos de classificação [29]. Alguns trabalhos utilizam técnicas automáticas para coleta de dados rotulados, tomando como ponto de partida termos conhecidos que expressam sentimentos positivos e negativos [43]. Em microtextos, algumas abordagens utilizam hashtags (#) conhecidas [44; 45] ou emoticons [31; 30; 46] para coletar dados rotulados automaticamente. No Twitter, as hashtags são utilizadas para criação de tópicos que são comentados também por outros usuários. Já os emoticons são caracteres que transmitem emoções. Em Li & Li [47], 87% dos tweets contendo emoticons possuem os mesmos sentimentos representados pelos emoticons no texto. Trabalhos que utilizam emoticons para treinamento dos classificadores têm apresentado excelentes re-

sultados de acurácia (acima de 80%).

Embora os emoticons apresentem uma ótima correlação com os sentimentos expressos nas mensagens dos tweets, estes estão presentes em menos de 10% dos tweets [48]. Assim, considerar uma análise de sentimentos que utiliza apenas os emoticons para determinar as polaridades dos sentimentos seria limitar as mensagens avaliadas, ignorando uma grande quantidade de tweets. O trabalho de Pak & Paroubek [31] utiliza a estratégia de emoticons para construir o conjunto de dados capaz de treinar um classificador Naïve-Bayes categorizando tweets opinativos em positivo ou negativo com base em N-gramas. Utilizando POS Taggers, os autores estudaram a distribuição das classes gramaticais contidas nos textos para diferenciar sentenças objetivas e sentenças subjetivas. A Figura 3.1 indica as frequências das classes gramaticais mais utilizadas em sentenças objetivas e subjetivas.

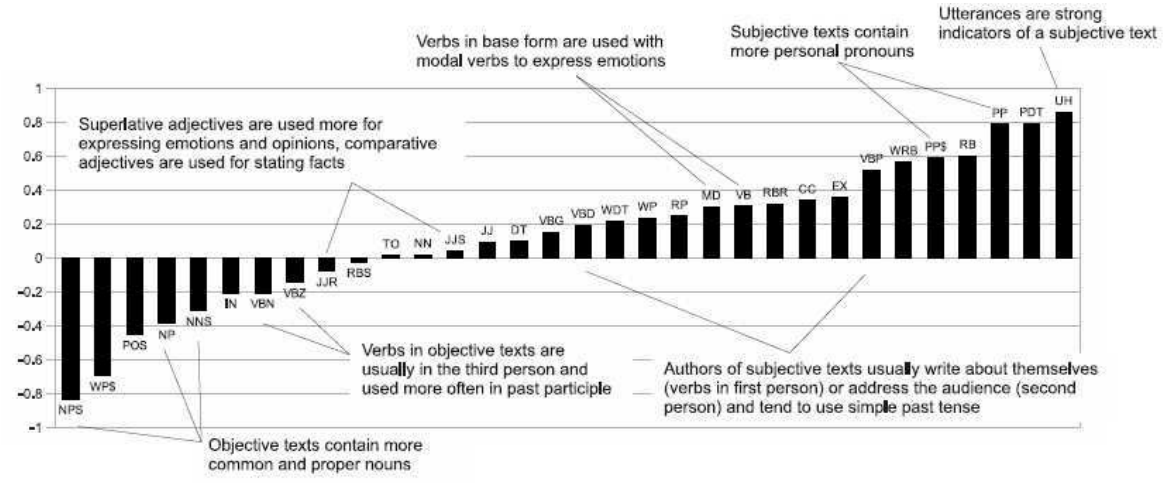


Figura 3.1: Classes gramaticais indicadoras de subjetividade e objetividade. Figura retirada de [31].

Para categorizar a polaridade do sentimento, Pak & Paroubek [31] utilizaram um classificador Naïve-Bayes. Outros trabalhos como o de Go et al. [30] e Read [46] relatam bons resultados utilizando o Naïve-Bayes para a classificação de sentimentos.

A tarefa de classificar o sentimento de um texto não é trivial, mesmo para seres humanos. O trabalho de Pang et al. [49] relata a dificuldade que as pessoas têm em distinguir palavras que expressam sentimentos positivos e negativos, ilustrando desta forma, a subjetividade na classificação do sentimento. O consenso na definição de polaridade do sentimento, quando realizado por seres humanos, dificilmente atinge 75%. Ao analisar comentários sobre filmes,

os autores avaliaram técnicas de aprendizado de máquina como, Naive Bayes, Entropia Máxima e SVM. Os autores utilizaram o Bag-Of-Words como característica dos classificadores e avaliaram ainda que os melhores resultados na classificação são obtidos com a utilização de unigrama. O classificador SVM obteve a acurácia de 82,9%.

No trabalho desenvolvido por Hu e Liu [6] é realizada uma análise de sentimentos em comentários sobre produtos em nível de aspectos. O trabalho dos autores caracteriza-se por: a) utilizar POS Tagging para identificar as características mais comentadas e as palavras de sentimento (adjetivos); b) aplicar uma abordagem semântica para classificação da polaridade do sentimento; e c) realizar a sumarização do sentimento tanto de forma global sobre produto quanto das suas características. Para definir a polaridade de uma sentença, uma lista inicial de palavras (*seed*) com a polaridade definida manualmente é utilizada e com o auxílio do WordNet [1] novas palavras de sentimento são adicionadas na lista dinamicamente através dos sinônimos (mesma polaridade) e antônimos (polaridade inversa), resultando assim em um dicionário específico de sentimentos para o domínio das revisões de produtos. O algoritmo também realiza um tratamento específico para palavras com sentidos de negação de sentença. Se a palavra de negação aparece perto de uma palavra de opinião (de acordo com um limite definido), a polaridade da palavra de opinião é invertida. A sumarização do sentimento do produto é realizada através da contagem do número de características positivas e negativas de cada revisão. Para testar o método, os autores extraíram 500 comentários de cinco produtos eletrônicos dos sites Amazon.com e Cnet.com. Na orientação semântica do sentimento das sentenças, o método obteve uma acurácia média de 0,84.

3.2 Trabalhos de Análise de Sentimentos Aplicados ao Idioma Português

Há poucos trabalhos na literatura que realizam análise de sentimentos com resultados utilizando um corpus em português. Os trabalhos de Chaves et al. [50], Sarmento et al. [51] e Tumitan & Becker [52] utilizam técnicas de análise léxica baseada em dicionários. Em Chaves et al. [50] é apresentado um algoritmo denominado de PIRPO (Polarity Recognizer

[1] um dicionário de palavras que contém sinônimos e antônimos. Mais informações em: http://wordnet.princeton.edu/

in Portuguese) que utiliza ontologias e lista de adjetivos polarizados (positivo, negativo e neutro), através do dicionários de sentimentos OpLexicon [53], para definir a orientação semântica dos textos analisados. Os resultados do PIRPO obtidos indicam uma F-Measure de apenas 0,32 no reconhecimento da polaridade. Em Tumitan & Becker [52], os autores analisam os sentimentos dos comentários sobre políticos realizados em jornais e correlacionam os sentimentos expressos com as pesquisas de intenção de votos realizadas por institutos. O algoritmo de identificação de polaridade utiliza o dicionário de sentimentos SentiLex-PT [54]. Os resultados iniciais (baseline) de Tumitan & Becker [52] indicam uma acurácia de 35,54%, mas para melhorar este índice, obtendo uma acurácia de 58,52%, os autores adicionam no dicionário de sentimentos novas palavras e expressões idiomáticas utilizas nos próprios comentários, tornando o dicionário dependente do contexto.

Em Nascimento et al. [55] são utilizadas técnicas de aprendizado de máquina e os classificadores de sentimento são utilizados para avaliar as reações das pessoas no Twitter em relação as notícias vinculadas na mídia. Os resultados alcançados em termos de acurácia variaram de 70% a 80% de acordo com o tipo de notícia e classificador utilizado. No entanto, no trabalho não é abordado mecanismos de sumarização do sentimento.

Em Gomide et al. [56] foi proposto um aplicação computacional para realizar análise de sentimento em tempo real no Twitter através do monitoramento de mensagens relacionadas a casos de dengue no Brasil. Um modelo de regressão linear foi construído para prever o número de de casos de dengue através da correlação de tweets postados com as estatísticas oficiais dos casos reportados. Os autores mostraram que o Twitter pode ser usado para prever surtos de epidemias de dengue realizando agrupamento espaço-temporal dos tweets relacionados a casos da doença no Brasil. A análise de sentimentos neste trabalho foi utilizada apenas para filtrar as mensagens que expressam sentimentos relacionados com a dengue.

3.3 Sumarização do Sentimento Espacial-Temporal

Dentre vários trabalhos publicados na área de análise de sentimentos, há dois recentes que, combinados, convergem para a proposta desta pesquisa, a saber: Bjørkelund et al. [13] e Dias [18]. O trabalho de Bjørkelund et al. [13] foca na análise de sentimentos aplicada aos comentários de viajantes para auxiliar na escolha de hotéis através da extração das caracte-

rísticas das regiões e hotéis visitados, tendo como principais contribuições a utilização do aspecto temporal para auxiliar a identificação de mudança de opinião e a visualização em mapas da sumarização da análise de sentimentos das regiões. Os comentários dos hotéis são coletados dos sites TripAdvisor e Booking.com, sendo estes já geocodificados com a localização geográfica do hotel. Para a classificação do sentimento os autores utilizaram duas abordagens de classificação: o SentiWordNet[2], para obter uma classificação binária, e o classificador Naive-Bayes, para obter graus de classificação contendo cinco categorias. A Figura 3.2 ilustra o protótipo desenvolvido que possibilita usuários realizar consultas sobre os hotéis de acordo com a análise de sentimentos realizadas sobre os comentários, identificando facilmente regiões geográficas que contêm os melhores ou piores hotéis. Além do mais, ao clicar no mapa em um determinado hotel, é possível visualizar o histórico do sentimento detectado sobre o mesmo. No trabalho de Bjørkelund et al. [13] o aspecto espacial é utilizado apenas para prover uma visualização geográfica no estado atual da análise do sentimento, não sendo possível visualizar através de mapas a mudança de opinião ao longo do tempo.

Figura 3.2: Protótipo da visualização em mapas da análise do sentimento. Figura adaptada de [13]

Já no trabalho de Dias [18] são utilizadas técnicas de Recuperação de Informações Geográficas (Geographical Information Retrieval - GIR) para georreferenciar documentos a partir de evidências textuais e realizar análise de sentimentos, gerando assim mapas temáticos através da sumarização dos resultados. Para validar as técnicas de geocodificação, documen-

[2]Framework para análise de sentimentos em textos no idioma inglês através de uma abordagem léxica [57]

tos georeferenciados da Wikipedia são utilizados. A detecção da polaridade do sentimento é realizada através de um modelo de classificação de textos fornecido pela ferramenta LingPipe[3] utilizando duas escalas de polaridade: uma de dois pontos (positivo ou negativo) e outra de cinco pontos, onde 1 significa um sentimento muito negativo e 5 corresponde ao sentimento muito positivo. Para validar as técnicas de detecção de polaridade, foram coletados comentários do website Yelp.com sobre diversas áreas, incluindo restaurantes, hotéis e diversas lojas próximas a algumas universidades dos Estados Unidos. Para a escala de dois pontos, o método obteve uma acurácia de 80% e para a escala de cinco pontos foi de 50%. A Figura 3.3 apresenta um mapa contendo a distribuição dos sentimentos realizadas com a escala de polaridade em cinco pontos sobre os documentos georeferenciados.

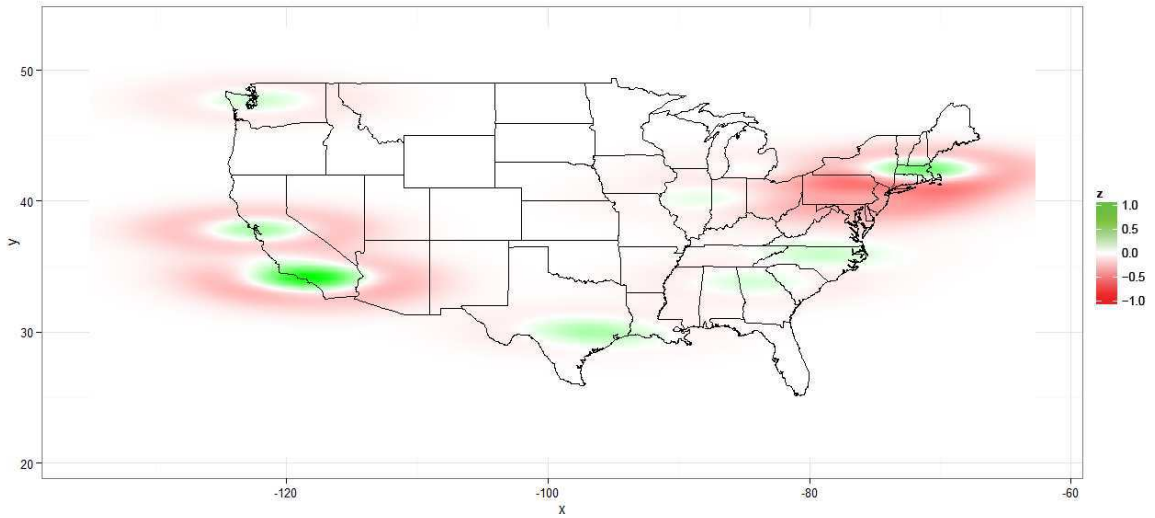


Figura 3.3: Distribuição espacial do sentimento em uma escala de 5 pontos. Figura retirada de [18]

3.4 Comparativo dos Trabalhos Relacionados

A Tabela 3.1 apresenta um comparativo contendo as principais características dos trabalhos relacionados. Para comparar os diversos trabalhos pesquisados, as características escolhidas foram:

- Abordagem da Análise de Sentimento: refere-se às técnicas utilizadas para a detecção

[3] http://alias-i.com/lingpipe/

do sentimento, conforme descritas na seção 2.1.4 do capítulo 2 ;

- Fonte de Dados: tipos de dados utilizados no estudo para realizar a análise de sentimento;
- Idioma: refere-se ao idioma da fonte de dados;
- Análise Temporal: indica se o trabalho realizava sumarização através de uma análise temporal dos dados;
- Análise Espacial: indica se o trabalho explorava a dimensão espacial dos dados, seja através da sumarização dos dados em mapas ou utilizando alguma técnica de GIR para georeferenciar os dados;
- Acurácia: indica a faixa dos resultados obtidos, em termos de acurácia, da técnica utilizada para detectar a polaridade do sentimento.

Tabela 3.1: Quadro Comparativo dos trabalhos relacionados

Autores	Abordagem de Análise de Sentimentos	Fonte de Dados	Idioma	Análise Temporal	Análise Espacial		Acurácia (%)
					Visualização	Geocoding	
Pang et al [50]	Aprendizado de Máquina (Naive-Bayes, Entropia Máxima, SVM)	Comentários sobre filmes	Inglês	não	não	não	77,0 - 82,9
Bjørkelund et al. [4]	SentiWordNet e Naive-Bayes	Comentários sobre hotéis	Inglês	sim	sim	não	68,0 - 84,0
Dias [8]	LingPipe	Comentários diversos (Yelp.com)	Inglês	não	sim	sim	50,0 – 80,0
Pak e Paroubek [7]	Naive-Bayes + Pos Tagger	Tweets	Inglês	não	não	não	75,0 – 85,0
Hu e Liu [2]	Pos Tagger + Análise Léxico (WordNet)	Comentários sobre Produtos	Inglês	não	não	não	84,0
Chaves et al.[51]	Análise Léxico	Comentários sobre hotéis	Português	não	não	não	F-Measure de 0,32
Sarmento et al.[52]	Análise Léxico + Regras Semânticas	Comentários sobre políticos	Português	não	não	não	77,0
Tumitan e Becker [53]	Análise Léxica (Lexicon-PT adaptado ao domínio)	Comentários sobre políticos	Português	sim	não	não	35,5 – 52,1
Nascimento et al. [56]	Naive-Bayes + N-Grama	Tweets	Português	não	não	não	70,0 – 80,0
Alves (2014)	Naive-Bayes e SVM	Tweets	Português	sim	sim	sim	65,0 – 88,0

3.5 Considerações do Capítulo

Diante deste cenário, esse trabalho difere, no que tange a classificação de sentimentos, das abordagens estudadas pela utilização de dois classificadores que elimina a necessidade de realizar PosTagger (Part-of-Speach) na identificação de um conteúdo opinativo. Assim, o primeiro classificador identifica se um conteúdo é subjetivo ou objetivo; e o segundo classificador identifica a polaridade (positiva ou negativa) do conteúdo já identificado como opinativo. Em relação à sumarização da opinião, este trabalho utiliza técnicas de GIR para geocodificar microtexto e explorar uma análise de sentimento espaço-temporal para visualizar os sentimentos detectados em mapas geográficos, incluindo a visualização de mudanças de sentimentos detectadas em diferentes períodos de tempo.

No próximo capítulo, será descrito em detalhes a abordagem de análise de sentimentos proposta neste trabalho.

Capítulo 4

Abordagem de Análise de Sentimento Espaço-Temporal

Este capítulo descreve a abordagem de análise de sentimentos desenvolvida neste trabalho que utiliza técnicas de Extração e Recuperação da Informação para sumarizar os sentimentos detectadas em microtextos. A abordagem de análise de sentimentos está dividida basicamente em quatro fases: extração dos dados, detecção da polaridade do sentimento em textos, detecção da região geográfica (geocoding) e a sumarização da opinião. Na seção 4.1, é apresentada a visão geral da abordagem de análise de sentimento proposta neste trabalho. Em seguida, na seção 4.2, é abordado o processo de extração de dados. Na seção 4.3, são apresentados os algoritmos implementados para a detecção da polaridade do sentimento. Na seção 4.4, é descrito o método utilizado no processo de georeferencimento dos microtextos. Finalmente, na seção 4.5, são descritas as formas de sumarização dos sentimentos realizadas através das dimensões espacial e temporal.

4.1 Visão Geral

A abordagem de análise de sentimento proposta neste trabalho é caracterizada pela utilização de técnicas de Recuperação da Informação (RI) para prover suporte à tomada de decisões através da mineração de informações contidas em microtextos, como os de microblogging e de redes sociais. Especificamente, este trabalho utiliza técnicas de Análise de Sentimentos para determinar a polaridade da opinião (positiva ou negativa) expressa em microtextos e

técnicas de Recuperação de Informação Geográfica (GIR) para inferir localizações geográficas a partir de evidências textuais. Os resultados obtidos através das técnicas mencionadas são utilizados para prover mecanismos de sumarização das informações, através de uma visualização espacial do sentimento em diversas regiões geográficas, incluindo também uma análise da variação do sentimento ao longo do tempo. O objetivo da sumarização é permitir uma melhor análise dos dados incluindo também o contexto geográfico.

A abordagem proposta neste trabalho é composta por quatro etapas, conforme ilustra a Figura 4.1. As seções 4.2, 4.3, 4.4 e 4.5 descrevem em detalhes essas etapas.

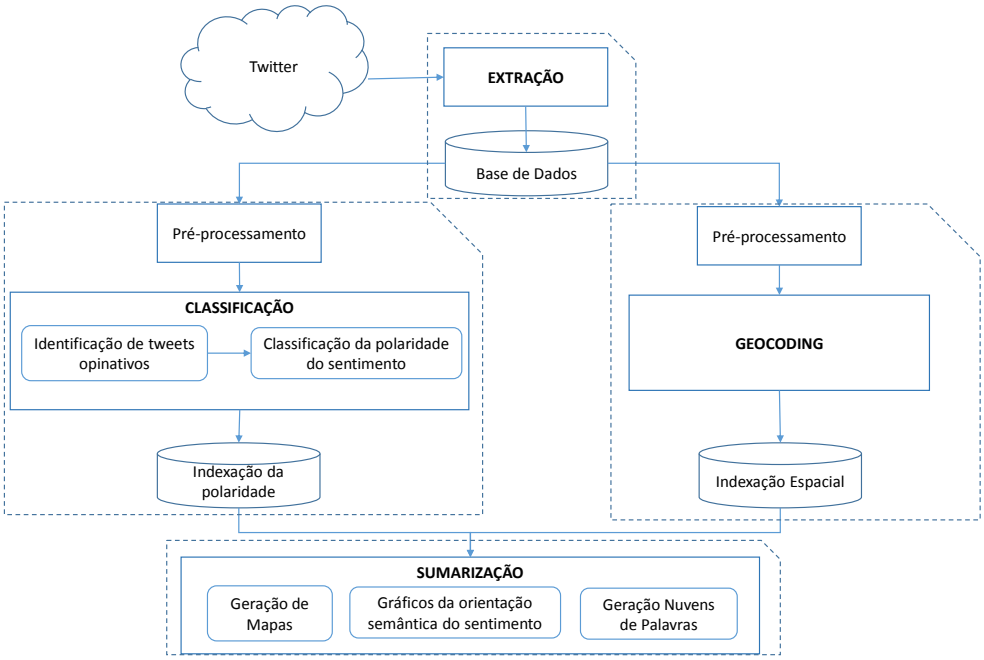


Figura 4.1: Visão geral da proposta de análise de sentimentos.

Após a coleta e armazenamento dos dados que serão analisados, a abordagem utiliza dois processos essenciais para obtenção dos resultados: detecção da polaridade do sentimento e geocodificação dos dados através das evidências textuais. Ambos utilizam técnicas de NLP para extrair e recuperar a informação em microtextos não estruturados, a exemplo dos tweets. O módulo de sumarização dos dados utiliza os resultados dos módulos de classificação do sentimento e geocodificação para prover mecanismos de visualização do sentimento detectado.

4.1.1 Definições

O escopo deste trabalho refere-se a microtextos relacionados a temas predefinidos. Assim, a análise de sentimento abordada é a nível de documento, não havendo necessidade de utilizar técnicas de Reconhecimento de Entidades Nomeadas (NER) para detectar a entidade avaliada. Portanto, considera-se que os sentimentos expressos nos microtextos, caso existam, sejam relativos aos temas predefinidos na coleta de dados.

Os microtextos explorados nesta pesquisa foram oriundos do microblogging Twitter[1], embora a proposta desta dissertação seja aplicável a outros tipos de microtextos, como os de redes sociais.

Eliminado o problema de NER, uma vez as entidades ou temas dos microtextos são previamente conhecidos, há basicamente dois problemas a serem resolvidos: detecção da polaridade do sentimento e geocodificação dos microtextos coletados.

O problema de detecção de polaridade do sentimento pode ser tratado como uma tarefa de categorização de textos. Mais formalmente, uma tarefa de classificação é encontrar uma função que aproxima a uma função de classificação $f : T \rightarrow C$ com $f(t_i) = c_j$ tal que $C = \{c_1, ..c_n\}$ representa um conjunto de n categorias predefinidas. A função f descreve como os textos são associados às classes, atribuindo um texto $t_i \in T$ para sua categoria $c_j \in C$. Neste trabalho, o conjunto T representa todos os textos coletados e $c_j \in C = \{positivo, negativo, neutro\}$ é a polaridade predominante (orientação semântica do sentimento) relacionada ao tweet t_i. Para definir a função de detecção de polaridade f, três algoritmos foram estudados e implementados conforme descrito na seção 4.3.

O processo de pré-processamento dos textos é realizado através de técnicas de NLP objetivando tratamento textual dos tweets para melhorar a detecção da polaridade. Formalmente, o pré-processamento pode ser definido como a aplicação de uma função $p : T \rightarrow T'$ tal que $p(t_i) = t'_i$ e $f(t'_i) = f(t_i) = c_j$, onde $t'_i \in T' = (t'_1, ..., t'_n)$ representa os tweets após realizar os tratamentos textuais.

Já o problema de geocodificação pode ser visto como um problema de encontrar uma função $g : T \rightarrow L$ tal que:

[1] www.twitter.com

$$g(t_i) = \begin{cases} l_m & \text{se } t_i \text{ apresenta referências geográficas} \\ 0 & \text{caso contrário,} \end{cases}$$

onde l_m representa a referência geográfica referenciada no texto t_i, com $l_m \in L = (l_1, .., l_m)$.

Desta forma, adicionando o contexto geográfico na definição formal de opinião[8] descrita pela equação 2.1, mencionada no Capítulo 2, seção 2.1, têm-se que, nesta abordagem de análise de sentimentos, a definição de opinião é representada através da sextupla abaixo:

$$O_{t_i} = (e_{t_i}, a_j^{t_i}, c_j, h_{t_i}, d, l_m), \tag{4.1}$$

onde:

- e_{t_i} representa a entidade relacionada ao microtexto t_i coletado, tal que $t_i \in T = (t_1, ..., t_n)$;
- $a_j^{t_i}$ é um aspecto/característica da entidade e_{t_i}, com $a_j^{t_i} \in A = (a_1^{t_i}, ..., a_m^{t_i})$, onde A representa o conjunto dos aspectos da entidade e_{t_i}. Este elemento é opcional, sendo utilizado apenas quando deseja-se um nível de detalhamento maior das entidades;
- c_j é a categoria que define a polaridade do sentimento sobre entidade e_{t_i}, ou o aspecto $a_j^{t_i}$ quando este for considerado. Esta categoria é obtida através da função f descrita acima;
- h_{t_i} é um emissor da mensagem t_i. Nesse contexto é o usuário do Twitter;
- d é a data e horário (instante) que a opinião foi expressa por h_{t_i}; e
- l_m representa a localização geográfica associada a t_i e que foi obtida através da função $g(t_i)$ descrita acima.

A equação 4.1 descreve todos os elementos necessários para uma análise de sentimentos que utiliza os aspectos espacial (l_m) e temporal (d). Neste trabalho, os elementos que representam os aspectos ($a_j^{t_i}$) e usuários h_{t_i} não foram utilizados nas implementações. A exploração destes elementos são encorajadas em trabalhos futuros para realização de uma análise de sentimentos mais detalhada.

4.2 Extração dos dados

O módulo de Extração dos dados é responsável por coletar os microtextos pesquisados armazenando-os de forma estruturada em um banco de dados para serem analisados pelos módulos de detecção de polaridade e geocodificação. O SGBD utilizado para armazenamento foi o PostgreSQL[2], com a extensão PostGIS[3] para realizar operações e processamentos espaciais.

Para a coleta dos dados, a fonte escolhida foi o Twitter. E, para tanto, foi desenvolvido um crawler na linguagem de programação Java, utilizando a biblioteca Twitter4J[4] que encapsula os serviços oferecidos na API[5] (Application Programming Interface) do Twitter, possibilitando a integração dos serviços com a aplicação Java.

A API do Twitter trabalha com quatro tipos de objetos[6]:

- Tweets (status updates): objeto básico do Twitter, sendo criado por usuários, podendo ser embarcado em sites, respondido, marcado como favorito e encaminhado (retweetado).

- Usuários: podem enviar tweets, seguir os tweets de outros usuários, criar listas de usuários, ser mencionados em outros tweets e monitorados por aplicações ou outros usuários.

- Entidades: são metadados e informações adicionais de contexto sobre os tweets, como por exemplo urls, hashtags e multimídia (fotos, videos, etc);

- Lugares: são localizações nomeadas que contêm coordenadas geográficas e podem ser associadas a tweets. Os lugares não necessariamente são de localizações geográficas dos usuários nos momentos de envios dos tweets, mas também podem ser outras localizações mencionadas nos tweets.

A API do Twitter oferece basicamente dois serviços: REST e Streaming. A API REST oferece as interfaces para a maioria das funcionalidades do Twitter e a API Streaming ofe-

[2]http://www.postgresql.org/
[3]http://postgis.net/
[4]http://twitter4j.org/
[5]https://dev.twitter.com/docs
[6]https://dev.twitter.com/docs/platform-objects

rece serviços em tempo real com o Twitter, mantendo uma conexão HTTP aberta com o microblog.

Neste trabalho, utilizou-se da API REST do Twitter, especificamente nos serviços de busca de tweets (GET search/tweets). No serviço de busca, há vários parâmetros que podem ser utilizados, mas o crawler desenvolvido utilizou os seguintes:

- Query (q): para obter tweets relacionados com as entidades de interesse para realizar a análise de sentimentos;
- Language(lang): para obter tweets apenas no idioma português (pt);
- Since e Until: especificar que os os tweets a serem recuperados foram criados entre as datas delimitadas: início (since) e fim (until);

Como o Twitter limita a quantidade de requisições da API dentro de uma janela de tempo, o crawler utilizou o serviço de forma autenticada, que garante maior quantidade de requisições, e quando esta janela é preenchida o crawler entra no modo de espera, respeitando o tempo limitado, para então continuar a busca dos tweets. Desta forma, todos os tweets que foram criados na data especificada são recuperados e armazenados no banco de dados.

Os resultados da busca na API Search do Twitter estão no formato JSON[7] (JavaScript Object Notation) contendo dados dos quatro tipos objetos do Twitter. No entanto, os principais dados utilizados neste trabalho foram: texto do tweet, data de criação, coordenadas de latitude e longitude, autor do tweet (usuário). Todos os dados obtidos no formato JSON também foram armazenados para realização de futuros trabalhos que visam explorar as entidades do Twitter e seus relacionamentos. A Figura 4.2 apresenta o esquema relacional do banco de dados utilizado para armazenar os dados coletados.

As entidades representadas pelas tabelas “tweets” e “tweets_json” são utilizadas para armazenamento das informações coletadas pelo crawler desenvolvido. As entidades “usuário”, “opinião” e “cidade” são utilizadas no sistema desenvolvido para formação do conjunto de dados de teste e treinamento, conforme será descrito na subseção 4.3.2 deste capítulo. Por fim, a entidade “classificação” é utilizada no processo de indexação da polaridade do sentimento.

[7] http://www.json.org

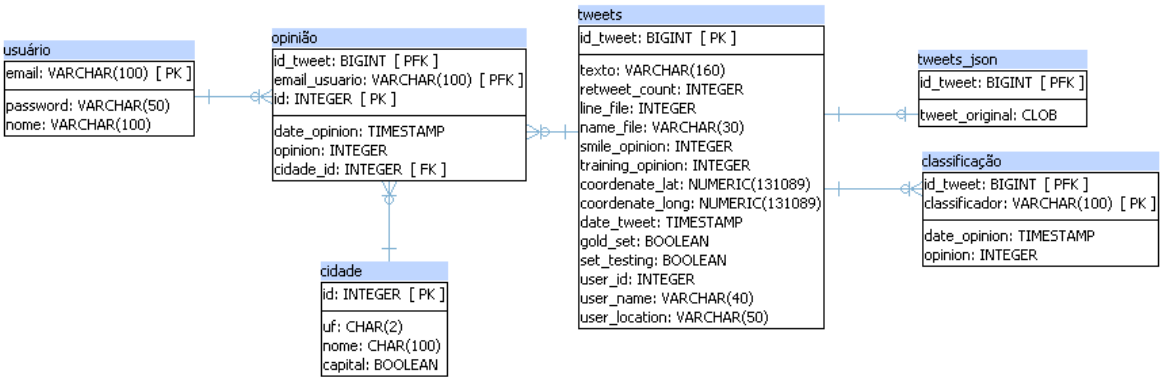


Figura 4.2: Esquema relacional do banco de dados.

4.3 Detecção da polaridade do sentimento

Nesta seção são descritas as técnicas implementadas para a detecção da polaridade do sentimento nos microtextos. O objetivo em avaliar e implementar técnicas diferentes de detecção de polaridade foi escolher a técnica que melhor detecta o sentimento em microtextos. As abordagens de análise de sentimento mais exploradas na literatura são: Análise Léxica, com base em dicionários de sentimentos, e Aprendizagem de Máquina Supervisionado. Ambas foram implementadas neste trabalho.

4.3.1 Análise Léxica

O aspecto central desta abordagem léxica é a utilização de dicionários de sentimentos, que são listas de palavras ou expressões de sentimento associadas a uma polaridade de sentimento, normalmente positiva ou negativa [8]. Embora os dicionários de palavras possam conter algumas informações adicionais, como por exemplo classe gramatical (adjetivos, advérbios, etc) e stemming (apenas o radical) das palavras, que auxiliam no processo de Part-Of-Speech, a informação essencial para a realização de análise de sentimento é a polaridade das palavras. A Tabela 4.1 apresenta um exemplo de alguns termos contidos em um dicionário de sentimento.

O algoritmo exibido no Código Fonte 4.1 ilustra a abordagem léxica implementada para a definição de polaridade. A ideia básica do algoritmo é obter as sentenças de um tweet (linha 4) e, através do processo de tokenização (linha 7), verificar a co-ocorrência das palavras do tweet com as palavras de sentimentos do dicionário (linha 11) e assim obter a orientação

Tabela 4.1: Exemplo de termos de um dicionário de sentimentos

Termo	**Classe Gramatical**	**polaridade**
abarrotado	adjetivo	-1
aberração	substantivo	-1
fértil	adjetivo	+1
fidedigno	adjetivo	+1
insinuante	adjetivo	0

semântica. Se na sentença apresentar uma palavra com sentido de negação, o sentido da polaridade do sentimento é invertido (linha 8).

Código Fonte 4.1: Orientação Semântica

```
1  getOrientacaoSemantica(tweet){
2      t = preprocessar(tweet);
3      orientacaoSemantica = 0;
4      for(sentenca: getSentencas(t)){
5        negacao = false;
6        sentSentenca =0;
7        for(token: getTokens(sentenca)){
8            if(token in ('não', 'nao')){
9              negacao = true;
10           }
11           if(token in dicionarioSentimento){
12             sentSentenca += getSentimentoDicionario(token);
13           }
14       }
15       if(negacao){
16         sentSentenca = -1*sentSentenca;
17       }
18       orientacaoSemantica += sentSentenca;
19     }
20     return orientacaoSemantica;
21 }
```

No processo de pré-processamento, termos que não contribuem para identificação da polaridade do sentimento são removidos, como stopwords, links e menções a usuários. Um

tratamento de HashTag é realizado, removendo o simbolo de "#" e desmembrando palavras compostas através da identificação de capitalização de letras. Por exemplo, "#foiMuitoBom" após o pré-processamento fica "foi Muito Bom".

A função "getSentimentoDicionario(token)" retorna um valor positivo (+1) caso a palavra (token) apresente um sentimento positivo, caso contrário, irá retornar um valor negativo (-1). Ao final do processamento, a variável o algoritmo resultará em uma valor positivo, neutro, ou negativo, indicando, desta forma, a orientação semântica do texto.

O tratamento de negação também foi realizado (linhas 16) invertendo o sentido da polaridade do sentimento obtido através do dicionário do sentimento. Por exemplo, na sentença "O jogo não foi bom", apenas a palavra "bom" é identificada com o sentido positivo (+) no dicionário de palavras, mas como uma palavra com o sentido de negação foi identificada, há uma inversão da polaridade do sentimento.

Segundo Bech e Tumitan [29], o método de co-ocorrência utilizado na abordagem léxica apresenta bons resultados quando o nível de análise textual é de granularidade pequena (análise em nível de sentença), a exemplo dos tweets que possuem no máximo 140 caracteres.

Quanto aos dicionários de sentimentos utilizados, no idioma português há apenas dois disponíveis: OpLexicon [53], para o português do Brasil e SentiLex-PT [54] para o português de Portugal. Nesta implementação optou-se por avaliar os dois dicionários.

4.3.2 Aprendizado de máquina

Em problemas de classificação de textos, a abordagem de aprendizagem de máquina supervisionado tem sido a mais explorada na literatura [58; 10]. Com a utilização de técnicas de aprendizagem de máquina supervisionada, faz-se necessário a formação de dois conjuntos disjuntos: treinamento e validação (teste). Então, dado dois conjuntos T_t e T_v, respectivamente de treinamento e validação, tem-se que $T_t \cap T_v = \oslash$. Assim, a classificação supervisionada começa com o conjunto de treinamento $T_v = (t_1, ..., t_n)$ com os textos já marcados com as categorias $c_j \in C = (c_1, ..., c_n)$ e a tarefa é determinar o modelo de classificação capaz de correlacionar corretamente um novo texto $t_w \in T_v$ à sua categoria, ou seja, $f : T \rightarrow C$ com $f(t_w) = c_j$. Neste caso, o conjunto de treinamento é utilizado para a construção de um classificador que aprenderá automaticamente as regras e características gerais dos documentos classificados. Já o conjunto de teste faz-se necessário para validar o classificador treinado.

Na tarefa de classificação, deve-se considerar as características (*features*) que o modelo de classificação deve observar nos dados de treinamento no processo de aprendizagem. Na classificação de textos, as porções (*tokens*) dos textos são extraídas e analisadas, e o classificador seleciona as características relevantes, representando-as na forma de um vetor de termos, ou bag of words, como é conhecido na literatura. Na Tabela 4.3, é ilustrado um exemplo de um vetor de termos binário de uma classificação de texto obtida pelo conjunto de treinamento da Tabela 4.2. O vetor de termos binário indica a ocorrência de termos nas classes, possibilitando a construção de modelos de classificação.

Tweet	Rótulo
Gostei do jogo do Brasil.	Positivo
Esse time do Brasil é muito ruim.	Negativo
Este não é um grande time.	Negativo

Tabela 4.2: Conjunto de Treinamento contendo a rotulação de tweets

gostei	jogo	Brasil	não	vai	grande	time	muito	ruim	classificacao
1	1	1	0	0	0	0	0	0	Positivo
0	0	1	0	0	0	0	1	1	Negativo
0	0	0	1	0	1	1	0	0	Negativo

Tabela 4.3: Exemplo de um vetor de termos

Em problemas de classificação de texto, especificamente de classificação de polaridade do sentimento, os tipos de características utilizados na literatura [8; 49; 20; 29] são:

Palavras de sentimento: utiliza dicionários de sentimentos para caracterizar de forma binária se as palavras de sentimentos estão presentes ou ausentes no texto;

Frequências de termos: utiliza-se n-gramas juntamente com a frequência absoluta ou relativa dos termos (TF-IDF);

Part-of-Speech: considera a classe morfológica (adjetivos, verbos, advérbios, etc) das palavras;

Neste trabalho, apenas a característica de frequência de termos foi utilizada nos classificadores implementados.

Construção dos conjuntos de dados para treinamento e validação

Um dos grandes problemas na utilização de aprendizado supervisionado para classificar a polaridade do sentimento é a necessidade de dados rotulados para treinamento do modelo que sejam apropriados para o domínio específico, uma vez que o desempenho do modelo treinado está fortemente relacionado com a qualidade e representatividade dos dados utilizados no treinamento.

Em trabalhos que tratam de revisões de produtos [26; 49] é comum a utilização da classificação realizada pelos usuários sobre os produtos na forma de notas ou estrelas, e nestes casos a rotulação das revisões para a formação do conjunto de treinamento pode ser realizada através desta classificação.

Para a criação dos conjuntos de dados (treinamento e validação) utilizados pelos classificadores através de técnicas de aprendizado supervisionado de máquina, neste trabalho foram utilizadas duas metodologias distintas: rotulação manual, através de auxílios de voluntários para classificação da polaridade, e rotulação automática através da presença de *emoticons* nos textos dos tweets.

Na rotulação manual, um sistema web, conforme ilustra a Figura 4.3, foi criado para possibilitar que usuários autenticados indicassem as polaridades dos sentimentos de uma amostra de tweets selecionados de forma randômica. Na indicação da polaridade, os usuários poderiam indicar as polaridades positiva, negativa e neutra. A polaridade neutra indica que através das evidências textuais (sintática e semântica) não há sentimentos expressos no tweet. O sentimento neutro pode indicar também a ausência de sentenças opinativas, tratando apenas de fatos ou acontecimentos. A utilização do voto majoritário é utilizada para tratar a detecção de sentimentos discrepantes entre as opiniões dos diversos usuários e, neste caso, somente a presença de um sentimento dominante valida a marcação do tweet no conjunto de dados. Desta forma, na hipótese de empate na indicação da polaridade do sentimento nos textos dos tweets, estes são descartados do conjunto da amostra (treinamento e validação).

Ainda nesta abordagem de rotulação manual, o usuário poderia indicar se o tweet continha referências a cidades do Brasil, criando assim um conjunto de dados também para a

validação da geocodificação do tweet que é tratado na seção 4.4.

Figura 4.3: Sistema para rotulação dos conjuntos de dados (treinamento e validação).

Na rotulação automática, o processo de identificação de polaridade baseia-se na análise de emoticons[8]. Nesta abordagem, assume-se que todas as palavras contidas no tweet têm os mesmos sentimentos expressado pelo emoticon. Assim, se um tweet apresenta um emoticon que transmite felicidade, por exemplo, a polaridade do tweet é considerada positiva. A utilização de emoticons tem sido adotada em outros trabalhos, como o de Pak & Paroubek [31], Read [46] e Go et al. [30]. A Tabela 4.4 apresenta os emoticons utilizados neste trabalho para rotulação automática na construção de dados para treinamento e validação.

Abordagens de Classificação

As abordagens de análise de sentimentos usualmente utilizam técnicas avançadas de NLP, como POS Tagging, para identificação de sentenças opinativas. Neste caso, o POS Tagging é utilizado para detectar mensagens subjetivas através da identificação da classes gramaticais das palavras contidas no texto. Em Pak e Pakoubek [31] foi verificado que em conteúdos subjetivos textuais há uma tendência de maior frequência de pronomes pessoais, adjetivos, adjetivos superlativos e comparativos, enquanto que textos objetivos ocorrem com mais frequência os nomes próprios e comuns. A tarefa de POS Tagging em textos informais,

[8]Um emoticon trata-se de uma sequência de caracteres tipográficos que transmite emoção, podendo expressar, por exemplo, sentimentos de felicidade ou tristeza.

Polaridade	emoticons
Negativo	D: D= D-: D^: D^= :(:[:{ :o(:o[:^(:^[:^{ =^(=^{ >=(>=[>={ >=(>:-{ >:-[>:-(>=^[>:-(:-[:-(=(=[={ =^[>:-=(>=[>=^(:'(:'[:'{ ='{ ='(='[=\ :\ =/ :/ =$ o.O O_o Oo :$:-{ >:-{ >=^{ :o{
Positivo	:) :] :} :o) :o] :o} :-] :-) :-} =) =] =} =^] =^) =^} :B :-D :-B :^D :^B =B =^B =^D :') :'] :'} =') ='] ='} <3 ^.^ ^-^ ^_^ ^^ :* =* :-* ;) ;] ;} :-p :-P :-b :^p :^P :^b =P =p \o\ /o/ :P :p :b =b =^p =^P =^b \o/

Tabela 4.4: Conjunto de Emoticons utilizados na rotulação automática.

como os tweets, não é uma tarefa fácil considerando problemas como: excesso de abreviações usadas nos textos ocasionado pela limitação de caracteres; palavras com letras repetidas para enfatizar termos; ausência de consoantes em palavras. Em textos escritos em português esses problemas são agravados, devido à complexidade gramatical inerente à linguagem.

Neste contexto, objetivando eliminar o uso de POS Tagging, esta pesquisa explorou o problema de análise de sentimento através de classificadores de texto, considerando apenas as palavras contidas no texto. Duas abordagens que utilizam técnicas de aprendizado de máquina supervisionado foram implementadas, conforme ilustra a Figura 4.4. Na primeira abordagem, nomeada de classificação simples, o modelo treinado é capaz de classificar as sentenças em três classes possíveis: positiva, negativa e neutra. Na segunda abordagem, o processo de classificação da polaridade é realizado com a utilização de múltiplos classificadores, mais precisamente com dois classificadores. O primeiro classificador é utilizado para classificar os textos em objetivos e subjetivos (opinativos), funcionando como um espécie de filtro para o próximo classificador. Já o segundo classificador foi treinado para classificar as polaridades das mensagens subjetivas em apenas duas classes: Positiva e Negativa.

Em detrimento de outros classificadores disponíveis, como KNN , Entropia Máxima, Árvores de Decisões e CRF, esta pesquisa utilizou os classificadores Naive Bayes e SVM

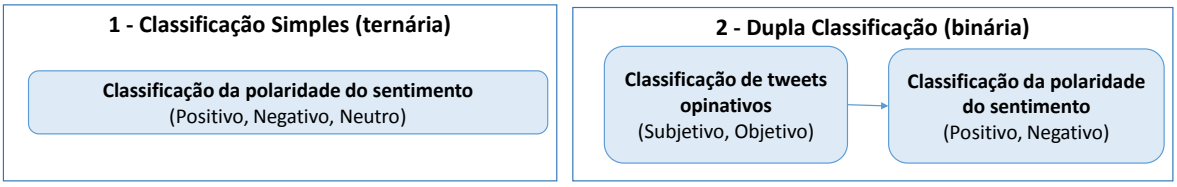


Figura 4.4: Abordagens de classifcadores.

para avaliar as abordagens de classificação do sentimento nos tweets, uma vez que estes classificadores são os mais explorados pela comunidade científica [28; 10; 31; 26; 49; 8].

Para construir os classificadores Naive Bayes e SVM, a API[9] do Weka [59] foi utilizada e incorporada na aplicação Java desenvolvida para avaliar os métodos de detecção de polaridade do sentimento. Na API do Weka, as classes que implementam os classificadores de Naive Bayes e SVM são, respectivamente, NaiveBayes.java e SMO.java. O classificador SMO (Otimização Minimal Sequencial) [60] é uma variante do SVM. Em ambos os classificadores, as frequências das palavras (Bag Of Word) foram utilizadas para representar as características dos textos no processo de treinamento dos modelos.

4.4 Identificação de Referências Geográficas

No processo de identificação das referências geográficas realizado neste trabalho, foi o utilizado o módulo do geoparsing do GeoSEn (Geographic Search Engine) [41]. A escolha de módulos do GeoSEn para realizar o processo de geocodificação neste trabalho deu-se em função de:

- Bons resultados no processo de geocodificação de documentos baseados na Web [61]:
  - 71% de referências válidas detectadas corretamente;
  - 92% de referências inválidas ignoradas corretamente;
  - 84% de acertos no processo de desambiguação;
- Base de dados geográfica de várias localizações do Brasil (Módulo do Geo Database)
- Aplicável em microtextos;

[9] http://weka.sourceforge.net/doc.dev/index.html

Para utilizar a API do GeoSEn com os microtextos que foram armazenados no banco de dados, foi necessário no processo de detecção de localização geográfica, converter cada a mensagem do tweet em uma página HTML (HyperText Markup Language), uma vez que o GeoSEn foi planejado para trabalhar com documentos Web. Então, após converter os microtextos para o formato adequado, os documentos Web são encaminhados para o módulo do Geoparser do GeoSEn, que é responsável pela detecção de termos geográficos dos textos analisados escritos em língua portuguesa. Nesta etapa, todos os locais candidatos são identificados e então encaminhados para a etapa seguinte, onde será aplicada a georeferência do texto. A Figura 4.5 ilustra um microtexto após a fase de Geoparsing, onde os termos candidatos à referências geográficas são detectados.

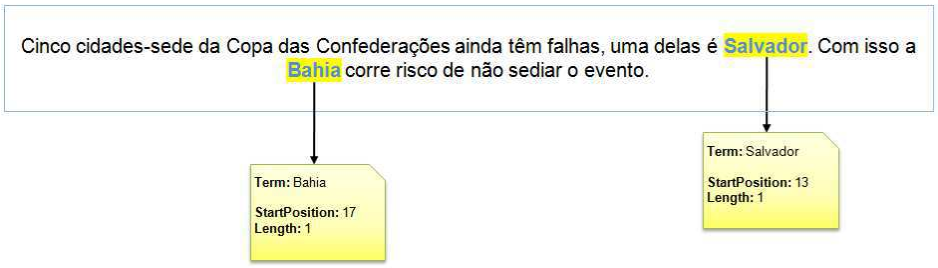


Figura 4.5: Exemplo do resultado do processo de Geoparsing sobre um microtexto (Figura retirada de [62]).

O Geoparser considera informações como a posição do termo em todo o texto e o seu comprimento, ou seja, a quantidade de palavras que formam o termo. Caso não seja detectado termos candidatos em um microtexto, este é considerado como microtexto que não possui referências geográficas, considerando as evidências textuais.

Após a identificação dos termos candidatos, estes são submetidos a uma avaliação de relevância para definição do escopo geográfico do microtexto. Nesta etapa, o Geo Scope Modeler do GeoSEn é utilizado, considerando a hierarquia geográfica de cidade $\mapsto$ microrregião $\mapsto$ mesorregião $\mapsto$ estado $\mapsto$ região para definir o escopo geográfico obtido através do cálculo da relevância que é realizado por meio dos níveis da hierarquia. A Figura 4.6 ilustra o resultado da etapa de Geocoding realizada sobre um microtexto.

Na Figura 4.6 é possível perceber que apenas um dos termos identificados na fase do Geoparser foi escolhido para georeferenciamento do microtexto. Isto ocorre devido as heu-

Figura 4.6: Exemplo do resultado do processo de Geocoding sobre um microtexto (Figura retirada de [62]).

rísticas do cálculo de escopo geográfico do GeoSEn, que considera o local geográfico mais preciso, que neste caso é o município.

4.5 Sumarização da opinião

Uma vez que os processos de detecção da polaridade de opinião e geocodificação dos textos dos tweets foram realizados, os resultados foram indexados no banco de dados para utilização na fase de sumarização da opinião, que é realizada através do módulo de sumarização dos dados. A análise visual é uma tarefa realizada com muita facilidade pelos humanos. A visualização é o meio pelo qual os seres humanos e os computadores cooperam com as suas capacidades distintas para obtenção de resultados mais eficazes [63]. Partindo deste princípio, a proposta de sumarização desta dissertação explora três opções, a saber:

1. **Análise Temporal do Sentimento:** associa os sentimentos positivos e negativos ao longo do tempo;

2. **Word Clouds do Sentimento:** gera nuvens de palavras através dos termos mais frequentemente citados nos tweets de acordo com o período e sentimento detectado;

3. **Visualização Espacial do Sentimento:** gera mapas associados aos sentimentos detectados, incluindo mapas de calor do sentimento;

Esta seção descreve em detalhes as abordagens de sumarização do sentimento implementadas nesta pesquisa.

4.5.1 Análise Temporal do Sentimento

A abordagem de análise temporal do sentimento, nesta pesquisa, é realizada através de gráficos de distribuição temporal, que agrupam a quantidade de mensagens por data, permitindo desta forma traçar o comportamento do sentimento geral detectado, auxiliando a detecção de momentos em que houve mudanças de opinião. A Figura 4.7 ilustra um gráfico distribuição temporal contendo o comportamento da quantidade de sentimentos positivos e negativos detectado no período analisado. A quantidade de sentimentos negativos foi plotada no semi-eixo negativo para evitar sobreposições com a do sentimentos positivos. Por exemplo, a partir do gráfico percebe-se que no dia 14/05/2014 foi detectado cerca de 3000 (três mil) tweets com polaridade negativa e cerca de 2000 (dois mil) tweets com polaridade positiva.

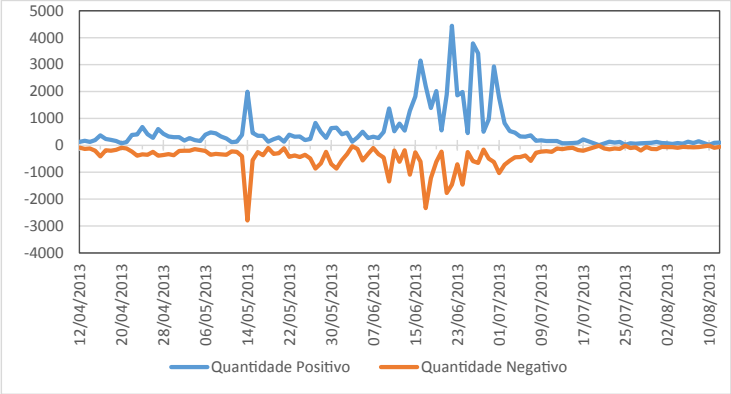


Figura 4.7: Análise do comportamento do sentimento detectado.

Outra alternativa de visualizar o comportamento da variação de tweets com polaridades positiva e negativa é através da fração (percentual) de tweets positivos e negativos por dia. Esta alternativa possibilita analisar a variação do sentimento, independente de quantidade de tweets com sentimentos, possibilitando a percepção da orientação do sentimento por dia em todo o período observado. Os gráficos contidos na Figura 4.8 ilustram as proporções de tweets positivos e negativos por dia. Por exemplo, percebe-se que no dia 14/05/2013, cerca de 40% de tweets com sentimentos positivos e 60% com sentimentos negativos.

Comparando com o gráfico da Figura 4.7, pode-se perceber que os gráficos da Figura 4.8

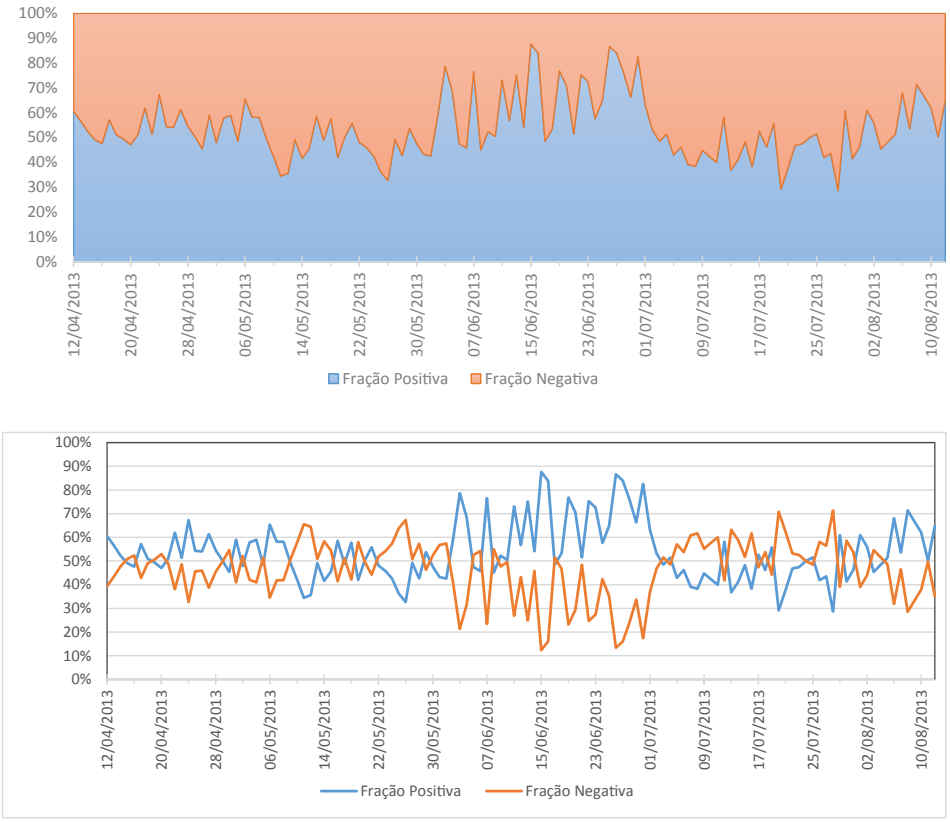


Figura 4.8: Análise da proporção de sentimentos positivos e negativos por dia.

auxilia na análise de sentimentos em períodos cuja a quantidade de tweets positivos e negativos seja pequena. Por exemplo, analisando o período dos dias 01/07/2013 à 10/08/2013, é possível afirmar que os gráficos da Figura 4.8 representa melhor a variação de sentimento ocorrida, possibilitando compreender o comportamento do sentimento neste período.

Uma forma de obter a orientação semântica geral do sentimento expresso nos microtextos é através da subtração do número de mensagens com sentimentos positivos pela quantidade de mensagens com sentimento negativo, ou seja, dado $Q_p = (p_1, ..p_n)$ e $Q_n = (q_1, ..q_n)$ com $p_i \in Q_p$ e $q_i \in Q_n$ representando, respectivamente a quantidade de mensagens com sentimentos positivos e negativos no tempo i, têm-se que a orientação semântica do sentimento geral pode ser obtido por $Q_{orientacao} = ((p1 - q1), ..(pn - qn))$. A Figura 4.9 apresenta um gráfico que ilustra a orientação semântica do sentimento geral.

A proposta deste trabalho foi criar uma forma interativa para que o tomador de decisão possa escolher qual o período que deseja analisar o comportamento do sentimento analisado, possibilitando inclusive exibir ou ocultar cada sentimento individualmente. A Figura 4.10

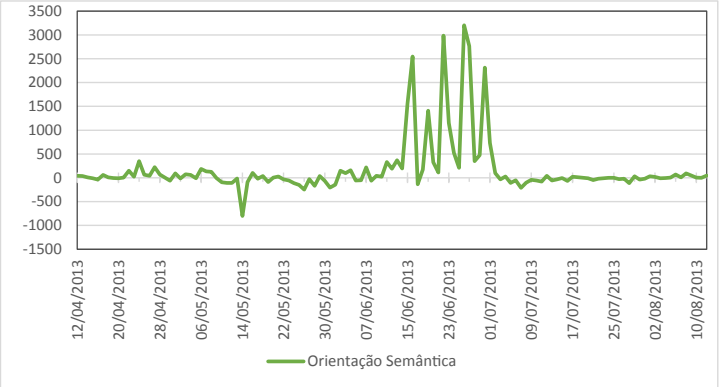
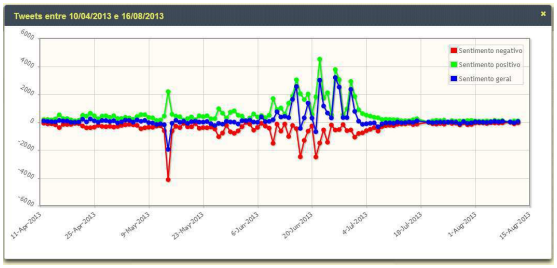
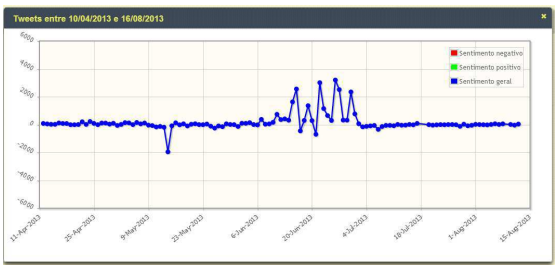


Figura 4.9: Análise do comportamento da orientação semântica do sentimento.

ilustra os resultados obtidos pela ferramenta desenvolvida que possibilita a visualização dos sentimentos detectados de forma interativa.

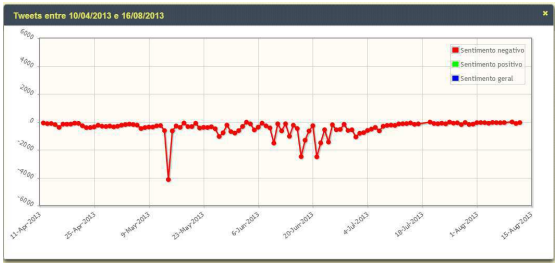


a) Distribuição Temporal dos Sentimentos



b) Apenas Orientação Semântica

c) Apenas Positivos



d) Apenas Negativos

Figura 4.10: Ferramenta interativa para análise temporal do sentimento detectado em micro-textos.

A Figura 4.11 ilustra o resultado de outro gráfico que possibilita a visualização dos sentimentos especificando o período de análise. A especificação de datas possibilita visualizar em detalhes o comportamento do sentimento em um período desejado e, em alguns casos, também é possível eliminar os efeitos causados pelos dados com outliers (quantidade de tweets atípicas), possibilitando a percepção de oscilações na tendência dos sentimentos com maior

eficácia, uma vez que os dados de outliers ofuscam visualmente essa percepção.

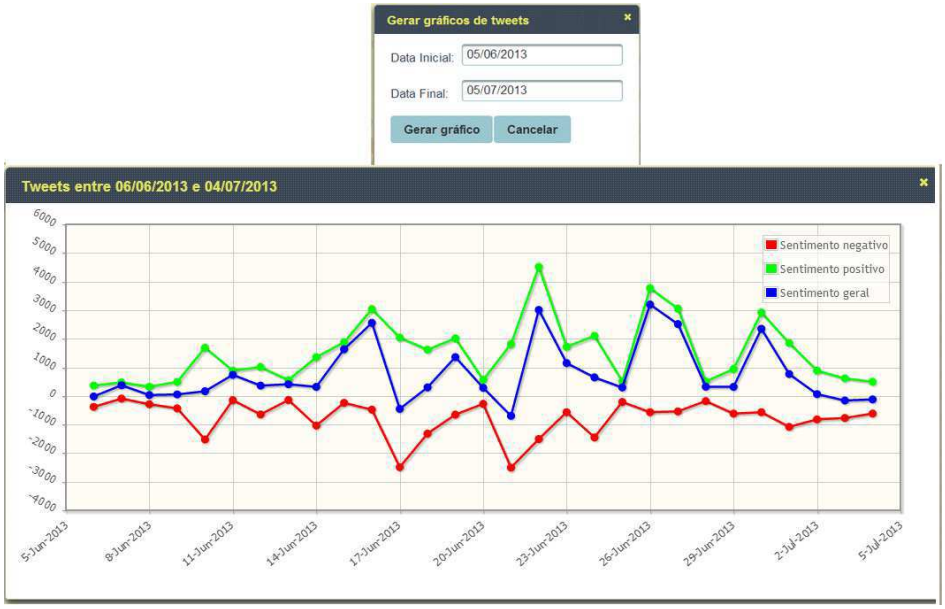


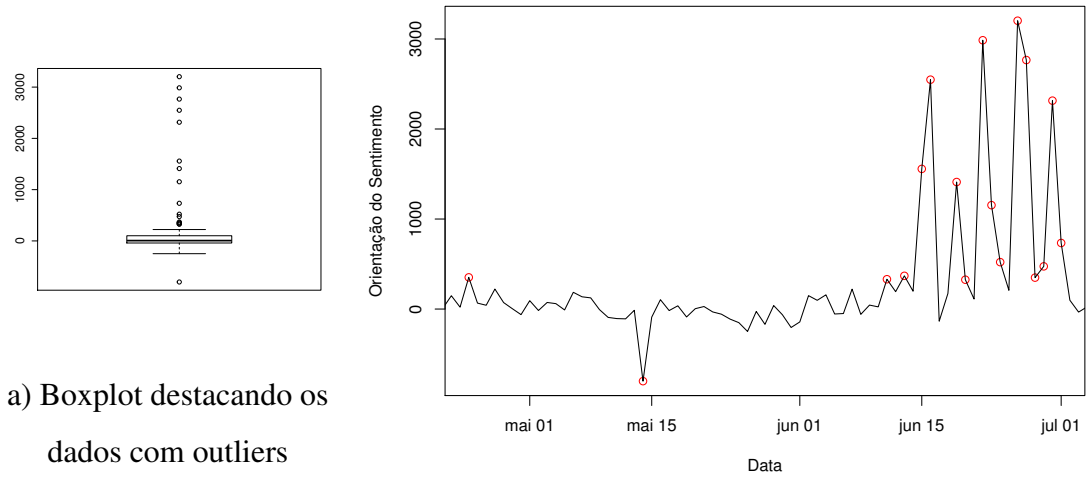
Figura 4.11: Análise do comportamento do sentimento obtida através da especificação do período.

4.5.2 Word Clouds do Sentimento

World cloud (Tag Word) é uma forma de visualização de dados textuais bastante popular, na qual ênfases são dadas em palavras chaves da fonte da informação. Normalmente, as palavras mais importantes são definidas através de critérios de ponderação, como por exemplo, frequências de ocorrência nos textos. A palavras são dispostas/arranjadas em uma "nuvem" de palavras, alternando o tamanho ou cor da fonte segundo o critério de importância. As nuvens de palavras possibilitam a compreensão gráfica de uma visão geral dos dados analisados.

Neste trabalho, a proposta de utilização de Word Cloud é através da exploração dos dados de outliers obtidos através da análise temporal, ou seja, permitir geração de word cloud nos dias em que as quantidade de tweets gerados estão fora dos quartis do gráfico de boxplot, conforme destacado na Figura 4.12.

A visualização das nuvens de palavras geradas em dias que houve quantidades atípicas de sentimentos detectados, sejam positivos ou negativos, pode possibilitar uma melhor compreensão dos acontecimentos que geraram os sentimentos, auxiliando o tomador de decisões



a) Boxplot destacando os dados com outliers

b) Orientação semântica do sentimento

Figura 4.12: Destaque dos outliers dos dados quantitativos relacionados a orientação semântica do sentimento.

sobre os aspectos que contribuíram para a expressão da opinião. A Figura 4.13 ilustra word clouds geradas em dias cujos os sentimentos predominantes foram, respectivamente, negativos(a) e positivos(b).

a) Polaridade Negativa do dia 14/05/2013

b) Polaridade Positiva do dia 21/06/2013

Figura 4.13: WordClouds geradas através de termos mais frequentes.

4.5.3 Visualização Espacial do Sentimento

Neste trabalho, a proposta de gerar mapas de sentimentos é através mapas de calor, conhecidos também como mapas de densidades. Um mapa de calor geográfico é um mapa de bits mostrando a densidade ou magnitude das informações analisadas relacionadas com as localidades geográficas. Com a geração de mapas de calor do sentimento é possível realizar uma análise geográfica do comportamento do sentimento detectado em diversas regiões geográficas.

Para a geração de mapa de calor, foi utilizado o GeoServer[10]. O GeoServer é um servidor gratuito baseado em Java que permite usuários visualizar e editar dados geoespaciais. Para geração dos mapas, utilizou-se uma camada de dados espaciais contendo a geografia de todos os estados brasileiros obtida junto ao Instituto Brasileiro de Geografia e Estatística (IBGE). Para gerar o efeito da distribuição da densidade (calor) no mapa, foi utilizada a extensão SLD[11] (Styled Layer Descriptor) do GeoServer que se encarrega de aplicar funções de renderização de acordo com os estilos de cores definidos e o conjunto de dados (sentimentos) georeferenciado. Uma transformação típica calcula a agregação a ser realizada de acordo com os dados de entrada, permitindo os efeitos visualização no mapa.

Para gerar mapas de calor, o GeoServer aplica uma função de renderização conhecida por "Vector-to-Raster" que gera uma superfície de "calor" no mapa através de um conjunto de pontos geográficos ponderados.

Neste trabalho, para gerar o conjunto de pontos ponderados no mapa, foram criados visões no banco de dados a partir da indexação do sentimento geográfico obtida no módulo gecoding, contendo cada localização geográfica detectada com os devidos quantitativos de sentimentos positivos e negativos detectados no módulo de classificação da polaridade do sentimento. Desta forma, a ponderação dos dados geográficos foram obtidas através de uma escala que define a proporção do sentimento positivo e negativo em cada localização detectada. Formalmente, a influência de cada localização geográfica no mapa de calor, é definida pela função $m : L \rightarrow E$ definida por:

$$m_{l_i} = \frac{quant^{l_i}_{pos} - quant^{l_i}_{neg}}{(quant^{l_i}_{pos} + quant^{l_i}_{neg})},$$

[10]GeoServer - OpenGeo. Mais informações em: http://geoserver.org/

[11]http://docs.geoserver.org/stable/en/user/styling/sld-extensions/index.html

onde $l_i \in L$ representa uma localização geográfica, $E = [-1, 1]$ é o peso da localização geográfica no mapa de calor obtido pelas quantidades de sentimentos positivos $quant_{pos}^{l_i}$ e negativos $quant_{neg}^{l_i}$ referente localização geográfica l_i.

Após obter o peso de cada região geográfica, o estilo para o SLD foi definido seguindo a escala de cores da Tabela 4.5.

Peso	Cor
-1	Vermelho
-0,5	Laranja
0	Amarelo (Branco)
0,5	Verde Claro
1,0	Verde Escuro

Tabela 4.5: Cores utilizadas no mapa de calor segundo o critério de ponderação

O estilo de cores (Tabela 4.5) é utilizado pela função de rasterização do GeoServer para que, através do peso de cada região geográfica e do agrupamento geográfico realizado no mapa, obter a escala de cores ilustrada pela Figura4.14.

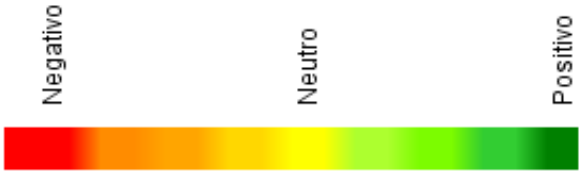


Figura 4.14: Legenda de cores utilizadas no mapa de calor de sentimentos

Assim, a obtenção do mapa de calor é realizada pelo GeoServer através de uma consulta no banco de dados que retorna todos os m_{l_i} referentes as localizações geográficas e o GeoServer através dos agrupamentos realizados gera os efeitos de densidades apropriados para visualização. As Figuras 4.15 e 4.16 ilustram mapas de calor gerados em períodos distintos em que foram realizadas a análise de sentimentos.

Na Figura 4.15, percebe-se que o período que a análise de sentimentos foi realizado a orientação semântica do sentimento, de forma geral, apresenta sentimentos positivos. Já na Figura 4.15, percebe-se que há algumas localizações geográficas cuja orientação semântica

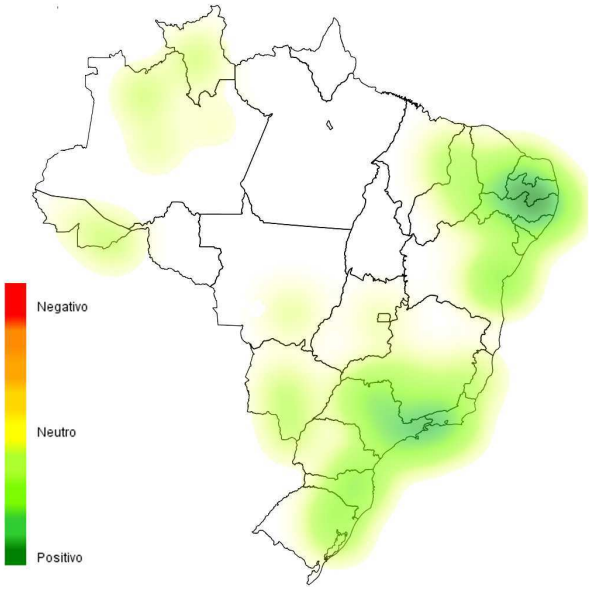


Figura 4.15: Mapas de calor: predominância de sentimentos positivos em todas as localizações

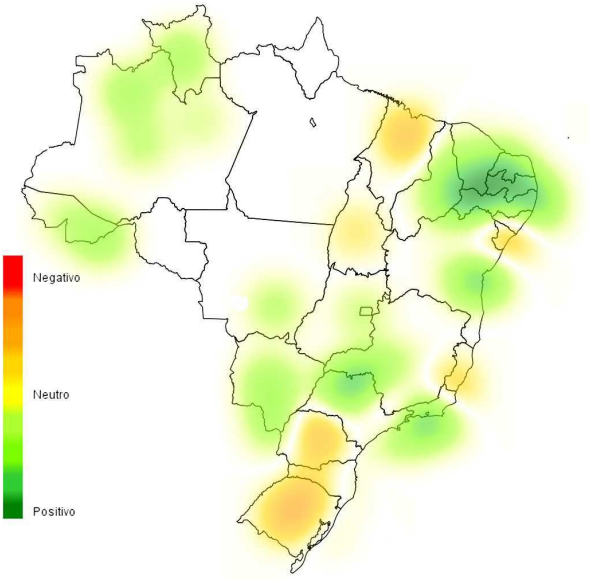


Figura 4.16: Mapas de calor: algumas localizações com a polaridade negativa

apresenta a polaridade negativa.

4.6 Considerações do Capítulo

Este capítulo apresentou a abordagem de análise de sentimentos que explora técnicas de sumarização do sentimento incluindo as dimensões espacial e temporal. Dentre as técnicas utilizadas nesta abordagem, como pré-requisitos para a sumarização do sentimento, destacam-se as técnicas analisadas para realizar a detecção da polaridade do sentimento e a aplicação de um módulo do GeoSEn para inferência de localizações geográficas em microtextos. Também foram abordadas as formas de sumarização dos sentimentos realizadas através das dimensões espacial e temporal, a saber: gráficos de distribuição temporal, word cloud e mapas de calor (densidade). No próximo capítulo, serão descritos os experimentos realizados através de um estudo de caso e as validações das técnicas de análise de sentimentos e georeferenciamento descritas neste capítulo.

Capítulo 5

Experimentos e Validação

Neste capítulo serão apresentados os resultados dos experimentos realizados com um caso de estudo sobre a Copa das Confederações de 2013 no Brasil, utilizando a abordagem de análise de sentimentos descrita no Capítulo 4. Na Seção 5, será realizada uma análise dos dados coletados. Em seguida, na Seção 5.1, os algoritmos implementados de detecção de sentimentos serão avaliados quanto a sua eficácia na detecção de polaridade e na Seção 5.2.3, o módulo de geoparsing do GeoSEn aplicado aos tweets coletados será avaliado. Por fim, na Seção 5.3.2 serão discutidos os resultados obtidos através da abordagem de análise de sentimentos proposta neste trabalho.

5.1 Coleta de Dados

Para execução dos experimentos relacionados com a proposta de análise de sentimentos desta dissertação, a API do Twitter[1] foi utilizada, conforme mencionado no Capítulo 4, para a criação do crawler que coletou aproximadamente 300.000 tweets no idioma português relacionados ao tema da Copa das Confederações da FIFA, realizada no Brasil em 2013. O crawler foi executado automaticamente todos os dias para coletar os tweets enviados no dia anterior ao do dia da execução cujos textos continham pelo menos um dos seguintes termos: #copa2014, #Brasil2014, Copa do Mundo de 2014, Copa das Confederações e #copadasconfederacoes. Embora existam termos que não tratam diretamente da Copa das Confederações, eles foram utilizados devido à forte relação com a Copa do Mundo, uma vez que a Copa das Confedera-

[1] https://dev.twitter.com/docs

ções é considerada também uma prévia da Copa do Mundo. Na Figura 5.1, é apresentado um gráfico contendo a quantidade de tweets obtidos através dos termos utilizado pelo crawler. Claramente, percebe-se que através do termo 'Copa das Confederações' foi obtido a maior quantidade de tweets.

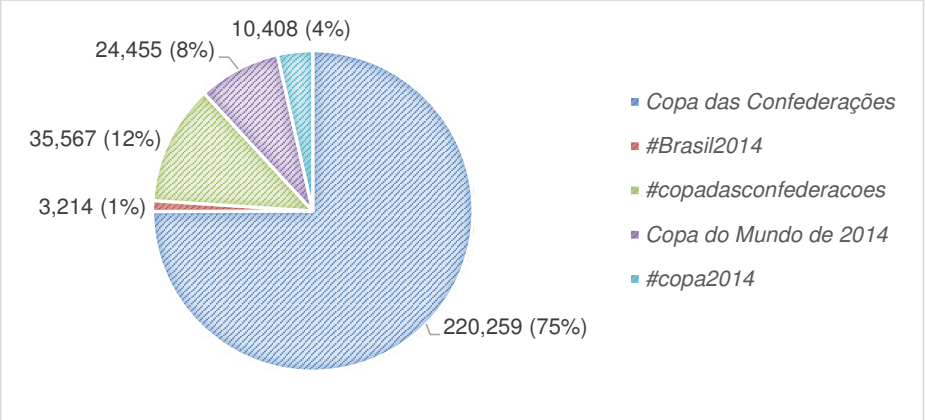


Figura 5.1: Números de tweets obtidos de acordo com os termos da coleta

O período de coleta de dados foi entre 12 de abril e 12 agosto de 2013, aproximadamente dois meses antes do início e dois meses após o término da competição, que ocorreu no Brasil no período entre 15 e 30 de junho de 2013. Compreender o período da coleta é importante para perceber o comportamento do sentimento expresso pelos brasileiros relacionadas ao tema. A Figura 5.2 ilustra a quantidade de tweets enviados diariamente durante o período coletado.

O pré-processamento dos textos foi realizado utilizando técnicas de NLP com o intuito de identificar e eliminar termos que não contribuem com a identificação da polaridade do sentimento tais como stopwords, links e menções a usuários. Para auxiliar este processo, foi utilizado o Apache OpenNLP[2] que é uma API Java utilizada em processamento de linguagem natural.

É possível observar (Figura 5.2) que no período da competição há uma maior quantidade de tweets enviados, como já era esperado. Percebe-se também que, no dia 14 de maio de 2013, houve uma quantidade atípica de tweets coletados, chegando a aproximadamente 17.000 tweets. A explicação para este acontecimento dá-se em função da divulgação da lista de jogadores convocados pela seleção brasileira de futebol e supõe-se que a maioria

[2]http://incubator.apache.org/opennlp

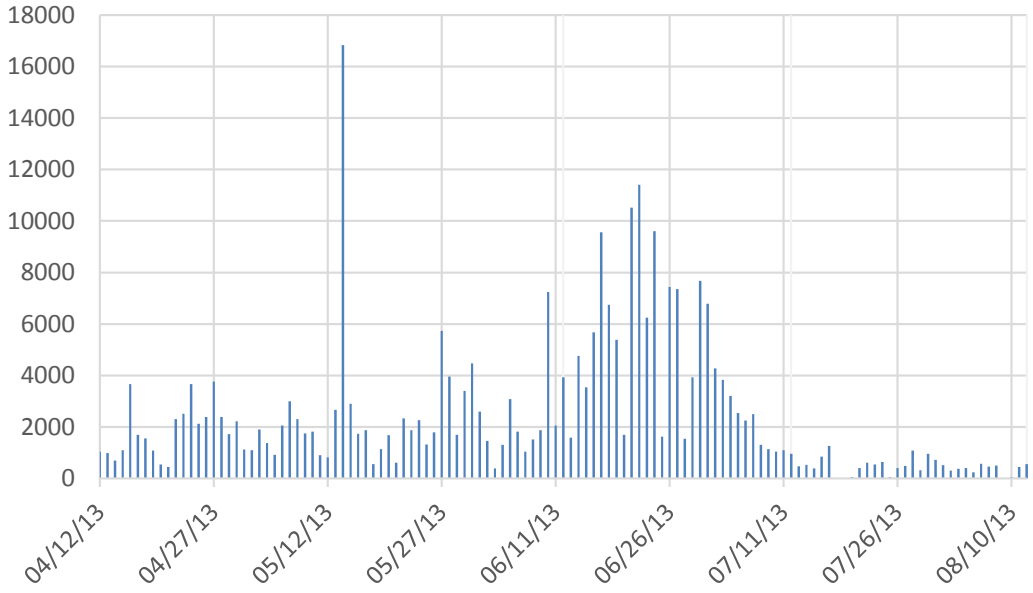


Figura 5.2: Número de tweets coletados por dia

destes tweets são opinativos, devendo retratar a opinião popular em relação aos jogadores selecionados para a formação do time na competição.

5.2 Avaliação dos algoritmos de detecção de opinião implementados

Nesta seção serão descritos os experimentos relacionados com a avaliação dos algoritmos de detecção de polaridade implementados nesta pesquisa. Inicialmente será abordada a metodologia utilizada para avaliação dos algoritmos; em seguida, será descrita a criação do conjunto de dados rotulados que foram utilizados para treinamentos e validação dos algoritmos de aprendizagem de máquina. Por fim, serão expostos os resultados dos algoritmos.

5.2.1 Metodologia da avaliação

As métricas utilizadas para avaliação dos resultados dos algoritmos de detecção de polaridade foram as tratadas no capítulo 2: acurácia, precisão, revocação e F-Measure. Para a avaliação dos algoritmos que utilizam técnicas de aprendizado de máquina supervisionado, uma fração dos tweets rotulados é reservada para treinar o modelo, não sendo utilizada para

a obtenção das métricas. A outra fração é utilizada para aplicar o classificador de sentimento e comparar os resultados com os rótulos marcados. A esses conjuntos denomina-se, respectivamente, dados de treinamento e dados de validação. Para avaliar a capacidade de generalização dos modelos de classificação, foi utilizado o método de validação cruzada $k - fold$, com $k = 10$, ou simplesmente $10 - fold$. O método $k - fold$ consiste em dividir o conjunto total de dados rotulados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir desta subdivisão, um subconjunto é utilizado para treinamento e os $k - 1$ restantes são utilizados para estimação das métricas utilizadas. Desta forma, o processo treina o modelo 10 vezes com dados de treinamentos distintos, avaliando as métricas com dados de testes também distintos a cada execução. Assim, as métricas são computadas utilizando a média das k execuções.

Já no processo de avaliação do algoritmo que utiliza uma abordagem baseada em dicionários (léxica) não há necessidade de realizar repetições, uma vez que o algoritmo é determinístico e não depende do conjunto de treinamento. Desta forma, as métricas foram computadas com apenas uma execução do algoritmo aplicado com o conjunto de tweets de testes.

5.2.2 Criação do Conjunto de Dados

Para avaliar os algoritmos implementados, fez-se necessário separar um conjunto para testes contendo os tweets já rotulados com as polaridades dos sentimentos. Para os treinamentos dos algoritmos de aprendizado de máquina supervisionado, fez-se necessário também separar um outro conjunto de tweets.

A rotulação do sentimento do tweet para obtenção do conjunto de dados, conforme mencionado no capítulo anterior, deu-se de duas formas: rotulação automática, através da utilização de emoticons, e rotulação manual, onde voluntários indicavam o sentimento nos tweets. Na rotulação automática, todos os tweets que apresentaram emoticons foram separados para formação do conjunto de dados. Já na rotulação manual, foram separados 1500 tweets de forma aleatória para que voluntários informassem os sentimentos dos tweets, utilizando o sistema online ilustrado na Figura 4.3 apresentado no capítulo 4.

As duas formas de rotulagem foram utilizadas, tanto para verificar se os modelos construídos através da rotulação automática podem ser considerados confiáveis, quanto para au-

Tabela 5.1: Número de tweets rotulados

Abordagem	Positivo	Negativo	Neutro	Total
Rotulação Automática	1.468	492	-	1960
Rotulação Manual	461	333	353	1227

mentar a quantidade de dados rotulados. A Tabela 5.1 apresenta o número de tweets rotulados com o sentimento segundo a forma de rotulação.

Observa-se, neste estudo de caso, que a utilização de emoticons que representam sentimentos positivos ou negativos (Tabela 4.4) em tweets estão presentes em menos de 1% do total de tweets coletados, sendo que dentre os tweets que apresentam emoticons, cerca de 75% apresentam emoticons relacionados com sentimentos positivos.

Na rotulação manual, dos 1500 tweets escolhidos aleatoriamente para formação do conjunto de dados, 80 tweets não foram rotulados por voluntários e 193 foram descartados por apresentarem divergência na identificação da polaridade do sentimento entre os voluntários. Desta forma, efetivamente foram considerados apenas 1227 tweets rotulados manualmente, e destes, considerando apenas aqueles tweets opinativos (positivos e negativos), cerca 58% são tweets que apresentaram sentimentos positivos. Comparado com a rotulação automática, percebe-se que a rotulação manual apresentou um melhor equilíbrio entre os quantitativos dos sentimentos detectados pelos voluntários.

5.2.3 Avaliação dos Algoritmos de Detecção de Polaridade

Algoritmos Baseados em Dicionários (Análise Léxica)

O algoritmo baseado em dicionários, exposto no capítulo anterior (4.1), foi o primeiro a ser implementado neste trabalho com o intuito de verificar a eficácia da abordagem nos tweets coletados. Como mencionado no capítulo anterior, há dois dicionários de sentimentos no idioma português e ambos foram utilizados no algoritmo implementado para comparação dos resultados. A Tabela 5.2 ilustra os resultados obtidos.

Como pode-se observar, os resultados alcançados não são satisfatórios, especialmente se considerar que o estado da arte em relação à detecção de polaridade do sentimento aponta para uma acurácia em torno de 90% [21]. Observando apenas a precisão de 91% na classe

Tabela 5.2: Resultados do algoritmo baseado no dicionário sentimentos

Dicionário	**Dados de Teste**	**Acurácia**	**Polaridade**	**Precisão**	**Revocação**	**F-Measure**
OpLexicon	Rotulação Manual	0.36	**Positive**	0.10	0.11	0.10
			Negative	0.20	0.25	0.22
			Neutra	0.91	0.90	0.91
			Média	0.40	0.42	0.41
SentiLex-PT	Rotulação Manual	0.37	**Positive**	0.29	0.40	0.33
			Negative	0.27	0.37	0.31
			Neutra	0.60	0.65	0.62
			Média	0.39	0.47	0.42

Neutra com a utilização do dicionário OpLexicon, inicialmente poderia representar um resultado muito bom. Mas a polaridade Neutra, conforme explicado no capítulo anterior, significa que não foi possível identificar o sentimento expresso no texto ou simplesmente que o texto não é opinativo (subjetivo). Este resultado acontece devido ao fato do algoritmo tentar localizar cada palavra do tweet no dicionário e, quando as palavras não estão contidas no dicionário, o algoritmo indica que o tweet é Neutro. Portanto, a precisão na classe neutra é mais alta com o dicionário OpLexicon porque este contém uma quantidade de termos maior que o SentiLex-PT.

Os resultados alcançados com a abordagem de dicionário (Léxica) são semelhantes a de outros trabalhos publicados que analisaram tweets no idioma Português utilizando a abordagem de dicionários, como o de Chaves et. al [50] e Becker e Tumitan [52]. No trabalho de Becker e Tumitan [52], o melhor resultado obtido, também utilizando a abordagem léxica, obteve uma acurácia em torno de 52%, mas que para isso, os autores adicionaram cerca de 268 novas palavras e expressões idiomáticas no dicionário de sentimentos utilizado (Lexicon-PT) após verificar os termos opinativos mais citados nos tweets analisados, tornando desta forma o detector de polaridade ainda mais dependente do contexto abordado.

Assim, diante dos resultados obtidos e considerando os de outros trabalhos que utilizam a abordagem de dicionários de sentimentos em tweets no idioma português, este trabalho explorou técnicas de aprendizado de máquina para realizar a classificação da polaridade do

sentimento.

Algoritmos Baseados em Aprendizado de Máquina

Como mencionado no capítulo anterior, para a classificação da polaridade dos tweets, duas abordagens foram utilizadas:

- Classificação Simples: modelo treinado é capaz de classificar os tweets com sentimento em positiva, negativa ou neutra.
- Dupla Classificação: o processo de classificação do sentimento é realizado em duas etapas, onde na primeira etapa os tweets são classificados em objetivos e subjetivos (opinativos) e na segunda etapa os tweets subjetivos são classificados nas polaridades positiva ou negativa.

A Tabela 5.3 contém os resultados obtidos pelos modelos de classificação SVM e Naive-Bayes quando aplicado na abordagem de classificação simples. No processo de validação cruzada $k - fold$ (Tabela 5.3), o conjunto de dados utilizados para treinamento e testes foi através da junção de todos os tweets rotulados. Ou seja, neste experimento o *DataSet* foi formado pela união dos dados obtidos através dos Emoticons (rotulação automática) com os da Rotulação Manual. A ideia inicial foi fornecer a maior quantidade de dados rotulados disponíveis para comparar os resultados dos classificadores SVM e Naive-Bayes. Os resultados indicam que, nesse processo de classificação, os dois classificadores apresentam resultados estatisticamente semelhantes, considerando um erro de 5%. O destaque está na precisão do modelo Naive-Bayes que apresenta um melhor resultado na polaridade positiva. Já em relação à métrica de revocação na classe positiva, o SVM apresenta melhores resultados.

Buscando especializar melhor o classificador de sentimentos para obter melhores resultados, a abordagem de classificação dupla que utiliza dois classificadores binários distintos, conforme já mencionado, foi também avaliada e os resultados constam na Tabela 5.4.

Ambos classificadores, Naive-Bayes e SVM, através da abordagem de classificação dupla, melhoraram bastante os resultados. No classificador SVM, a acurácia foi de 62,7% e 80,5%, utilizando as abordagens de classificação simples e dupla, respectivamente. Já no classificador Naive-Bayes a acurácia observada foi de 61,0% e 77,7%, respectivamente.

Tabela 5.3: Comparação dos Classificadores de Sentimentos - Classificação Simples

Classificador	**DataSet)**	**Acurácia**	**Polaridade**	**Precisão**	**Revocação**	**F-Measure**
SVM - Simples	Emoticons + Rotulação Manual	0.627	**Positiva**	0.665	0.822	0.735
			Negativa	0.651	0.542	0.592
			Neutra	0.431	0.286	0.344
			Média Ponderada	0.610	0.628	0.610
Naive-Bayes - Simples	Emoticons + Rotulação Manual	0.610	**Positiva**	0.804	0.607	0.692
			Negativa	0.535	0.706	0.608
			Neutra	0.427	0.498	0.460
			Média Ponderada	0.647	0.610	0.618

Tabela 5.4: Comparação dos Classificadores de Sentimentos - Classificação Dupla

Classificador	**DataSet)**	**Acurácia**	**Polaridade**	**Precisão**	**Revocação**	**F-Measure**
SVM - Classificação Dupla	Emoticons + Rotulação Manual	0.805	**Positive**	0.839	0.873	0.856
			Negative	0.715	0.657	0.685
			Média Ponderada	0.799	0.802	0.800
Naive-Bayes - Classificação Dupla	Emoticons + Rotulação Manual	0.777	**Positive**	0.91	0.742	0.817
			Negative	0.616	0.849	0.714
			Média Ponderada	0.813	0.777	0.783

Embora o resultado com o classificador SVM seja ligeiramente melhor que o classificador Naive-Bayes, os resultados de ambos são considerados satisfatórios para Tweets no idioma português, e representam os melhores resultados dentre os trabalhos pesquisados na literatura.

Já para avaliar a influência do conjunto de dados, a Tabela 5.5 apresenta os resultados dos classificadores utilizando diferentes conjuntos de dados para treinamento e testes. Observa-se que, de fato, a utilização de emoticons para rotulação automática dos sentimentos nos tweets possibilita uma forma eficiente e rápida para construção de conjunto de dados para treinar os modelos de classificação de sentimentos, conforme indicado em Pak e Paroubek [31].

Tabela 5.5: Comparação do Conjunto de Dados (Treinamento e Testes.)

Classificador	**DataSe)**	**Acurácia**	**Polaridade**	**Precisão**	**Revocação**	**F-Measure**
SVM - Classificação Dupla	Emoticons	0.870	**Positiva**	0.953	0.847	0.897
			Negativa	0.748	0.916	0.824
	Rotulação Manual	0.656	**Positiva**	0.762	0.716	0.738
			Negativa	0.469	0.529	0.497
Naive-Bayes Classificação Dupla	Emoticons	0.727	**Positiva**	0.820	0.765	0.791
			Negativa	0.569	0.649	0.606
	Rotulação Manual	0.650	**Positive**	0.805	0.636	0.710
			Negative	0.472	0.678	0.556

Com o objetivo de melhorar os resultados dos classificadores, alguns tratamentos textuais dos tweets foram realizados, como por exemplo:

- Filtragem: remoção de URL's, nomes de usuários do Twitter (inicia com @) e palavras especiais do Twitter (RT e via);
- Remoção de stopwords;
- Tratamento dos termos compostos que estejam com Hashtags. Normalmente é realizado a separação dos termos seguindo a orientação de capitalização das letras. Por

exemplo: #BomDemais tratado fica “Bom Demais” (adiciona-se um espaço entre as palavras);

- Remoção de letras repetidas: Por exemplo, a frase “O time do Brasil é muito boooo-ommmm” é substituído por “O time do Brasil é muito bom”.

No entanto, os resultados dos classificadores não apresentaram melhorias significativas. O melhor resultado foi um aumento na acurácia de 0,2% para o tratamento de Hash Tag. Acredita-se que um estudo mais detalhado através de uma conjunto de dados maior para testes no processamento dos textos possa contribuir para melhorar os resultados obtidos.

5.3 Avaliação da Detecção de Referências Geográficas em Microtextos

Embora o módulos de detecção de lugares (geoparsing) e modelagem do escopo geográfico dos documentos (georeferencing) do GeoSEn já tenham sido avaliados quanto a eficiência na inferência geográfica em documentos baseados na WEB [41], faz-se necessário verificar os resultados no contexto deste trabalho, ou seja, quando os módulos são aplicados a microtextos, como os dos tweets coletados. A validação da sumarização espacial depende diretamente da eficiência do processo que envolve a detecção das localidades geográficas nos microtextos.

5.3.1 Construção e Análise do Conjunto de Tweets Georeferenciados

Embora os tweets contenham um atributo que possibilita ao o usuário georeferenciar suas mensagens, este é opcional e não há garantias que o texto abordado na mensagem tenha relação com o conteúdo do tweet. Por ser opcional, são poucos usuários que utilizam os atributos de referências geográficas. Dos cerca de 300.000 tweets coletados, apenas 5245 apresentaram esta informação, o que representa menos que 2% do total de tweets. A premissa deste trabalho é que é possível aumentar a quantidade de tweets, incluindo o correto georeferenciamento das mensagens através técnicas de GIR.

No sistema de rotulagem manual da polaridade de tweets manual desenvolvido, conforme ilustrado no capítulo anterior, o usuário também poderia rotular as mensagens do tweet no

que tange a localização geográfica. No entanto, da amostra dos 1500 tweets escolhidos aleatoriamente, dos quais 227 foram descartados do conjunto de testes porque os voluntários não emitiram opinião sobre localização geográfica, apenas uma quantidade de 137 tweets continham evidências textuais suficientes para serem georeferenciados. Embora a quantidade de tweets georeferenciados da amostra seja pequena, as inferências estatísticas podem ser realizadas, desde que se considere um erro amostral de 5% com um nível de confiança de 95%, dado que o tamanho da amostra é bastante representativo, considerando os parâmetros do erro amostral e do nível de confiança com o tamanho da população (total de tweets) e o fato de que a amostra foi obtida de forma aleatória.

A Figura 5.3 ilustra a quantidade de tweets da amostra georeferenciada por cidade. Observa-se que as cidades sede dos jogos da Copa das Confederações de 2013 são aquelas cujos tweets apresentaram maior quantidade georeferenciados, conforme ilustra Tabela 5.6 [3]. Embora em Brasília tenha ocorrido apenas um jogo, na amostra de tweets com referências geográficas detectadas, ela é a segunda cidade com mais tweets georeferenciados. Isto deve-se ao fato de que foi em Brasília onde ocorreu a abertura da competição. Já com relação ao Rio de Janeiro, que tem quase 50% da amostra, a explicação se dá pelo fato de que ela foi a cidade onde ocorreu o último jogo da competição, que define o time campeão. Esta análise é importante de ser realizada, porque dado que a amostra é significativa, o resultado do processo de identificação das referências geográficas realizado de forma automática deve refletir o comportamento desta análise.

Tabela 5.6: Quantidade de Jogos nas Cidades Sede.

Cidade	**Quantidade de Jogos**
Rio de Janeiro - RJ	3
Brasília - DF	1
Belo Horizonte - MG	3
Fortaleza - CE	3
Salvador - BA	3
Recife - PE	3

[3]Dados obtido da WikiPedia. Disponível em http://pt.wikipedia.org/wiki/Copa_das_Confederações_FIFA_de_2013. Acessado em 25/07/2014.

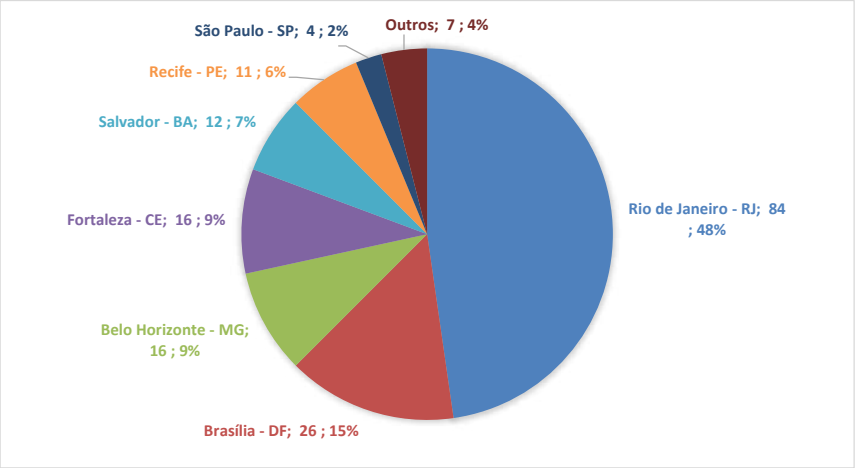


Figura 5.3: Quantidade de tweets georeferenciados da amostra por cidade

5.3.2 Validação da Técnica de Detecção de Referências Geográficas

Considerando a amostra validada pelos voluntários com relação as referências geográficas contidas nos tweets e o resultado do processo de identificação de localizações geográficas realizado pelos módulos do GeoSEN, a Tabela 5.7 apresenta a Matriz de Confusão pela qual as métricas de Acurácia, Precisão, Revocação e F-Measure são obtidas. Na Tabela 5.8 constam os resultados dos quais permitem uma análise da eficiência do processo de identificação de regiões geográficas.

Tabela 5.7: Matriz de Confusão da Análise das Referências Geográficas nos Tweets.

	Localização Geográfica Rotulada Manualmente		
		Sim	Não
Localização Geográfica Processada	Sim	72	6
	Não	65	131

Na Matriz de Confusão (Tabela 5.7), dado que foram processados 274 tweets, têm-se que:

- houve 72 tweets com a localização geográfica identificada corretamente(verdadeiros

positivos);

- foram detectados 6 tweets com a localização geográfica identificada incorretamente (falsos positivos);
- foram detectados 131 tweets sem identificação de referências geográficas por de fato não haver evidências textuais (verdadeiros negativos); e
- houve 65 tweets com a localização geográfica que não foi identificada (falso negativos).

Tabela 5.8: Métricas do Resultado da Identificação de Referências Geográficas

Resultados			
Acurácia	**Precisão**	**Revocação**	**F-Measure**
74,08%	92,30%	52,55%	66,97%

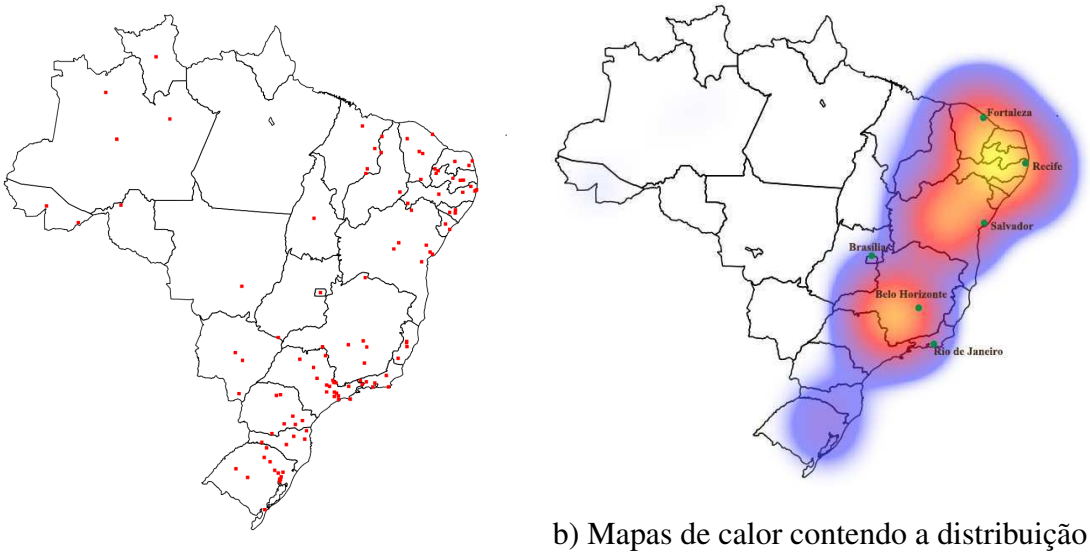
Das métricas analisadas, conclui-se que os módulos de detecção de referências geográficas do GeoSEn apresentam como resultado um baixo índice de revocação, ou seja, existem 47,45% de tweets com referências geográficas (documentos relevantes) que não foram identificadas as referências (documentos recuperados). Tal resultado já era esperado, uma vez que o escopo geográfico considerado no GeoSEn segue a hierarquia de país até município, desconsiderando outras referências geográficas dentro dos municípios, como os Estádios de Futebol e os Aeroportos (ambos bastantes citados neste contexto). Por outro lado, a precisão apresenta um excelente resultado, ou seja, a quantidade de referências válidas detectadas apresenta um índice de acerto de 92,3%.

5.4 Análise do Sentimento Espaço-Temporal

Nesta seção, os dados numéricos, já validados nas seções anteriores deste capítulo, serão analisados com os resultados gráficos obtidos através da análise de sentimento espaço-temporal.

5.4.1 Distribuição Espacial dos Tweets Coletados

Após a realização do processo de detecção de referências geográficas aplicado a todo o conjunto de tweets coletados (cerca de 300.000), foram realizadas inferências geográficas em aproximadamente 7560 tweets, por intermédio dos módulos de geocodificação do GeoSEn. A Figura 5.4(a) ilustra a distribuição espacial dos tweets que tiveram localizações geográficas detectadas. No entanto, utilizando apenas este mapa, não possível conhecer os quantitativos dos tweets em cada localidade. Mas, com o auxílio do mapa de calor 5.4(b), o qual evidencia a densidade de tweets enviados, pode-se perceber as localizações em que mais houve detecção de localizações nos tweets coletados.



a) Localidades geocodificadas em tweets

b) Mapas de calor contendo a distribuição de tweets georeferenciados

Figura 5.4: Visualização do quantitativo dos tweets georeferenciados

Desta forma, utilizando o mapa de calor da Figura 5.4(b), pode-se confirmar a inferência estatística realizada na seção 5.3 através da amostra, que indicava uma concentração de tweets com as referências geográficas nas cidades sede da Copa das Confederações. Esta informação têm índice de precisão em torno de 92%, conforme analisado na subseção 5.3.2 deste capítulo.

Comparando ainda os dois mapas da Figura 5.4, pode-se observar que as localizações geográficas detectadas nos tweets na região norte (Amazonas, Acre, Rondônia e Roraima) não evidenciam-se no mapa de calor. Isto ocorre devido ao fato de que a quantidade de tweets com essas referências geográficas é insignificante, comparado com a região nordeste

e sudeste.

5.4.2 Análise Espaço-Temporal

A proposta desta dissertação é a realização de uma abordagem de análise de sentimento espaço-temporal que possibilita ao tomador de decisão, compreender a distribuição do sentimento detectado em microtextos considerando os aspectos espacial e temporal. Nesta abordagem, é possível escolher o período sobre o qual pretende-se visualizar os mapas de calor, tornando viável a identificação de comportamentos relacionados aos sentimentos nas diversas localizações geográficas. Para exemplificar, a análise espaço-temporal, quatro períodos de tempo foram escolhidos para análise, conforme ilustra a Figura 5.5.

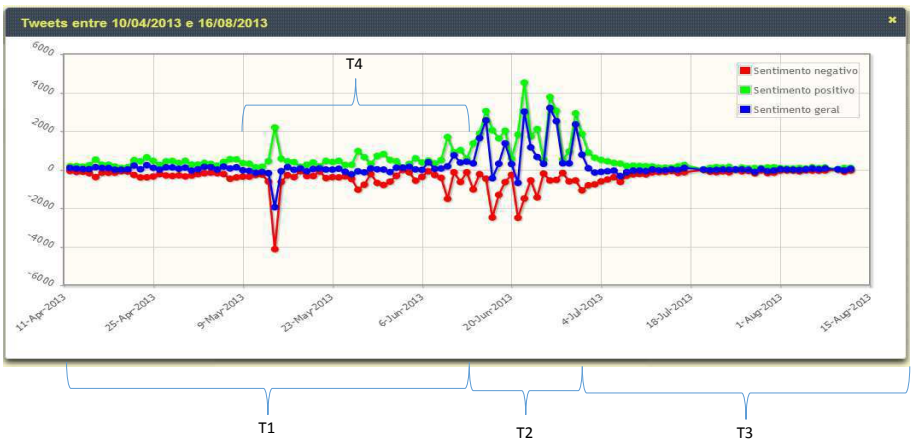


Figura 5.5: Análise do Sentimento Espaço-Temporal: Períodos Analisados

Os períodos $T1$, $T2$ e $T3$ foram escolhidos por corresponderem, respectivamente, às fases que antecederam a competição, que a compreendiam e que sucederam à mesma. Já o período $T4$ foi escolhido por dois motivos: 1) compreender um dia em que houve muitos tweets enviados relacionados convocação dos jogadores da seleção do Brasil; 2) compreender um período com maior números de tweets com sentimentos negativos. Assim, analisar as polaridades dos sentimentos nestes períodos, considerando o fator espacial, possibilita a identificação do comportamento das pessoas no tempo e espaço. A Figura 5.6 ilustra o resultado da análise de sentimento espaço-temporal realizada sobre os períodos de tempo $T1$, $T2$, $T3$ e $T4$.

A Figura 5.6 (a) ilustra que, no período que antecedeu a competição, o sentimento ex-

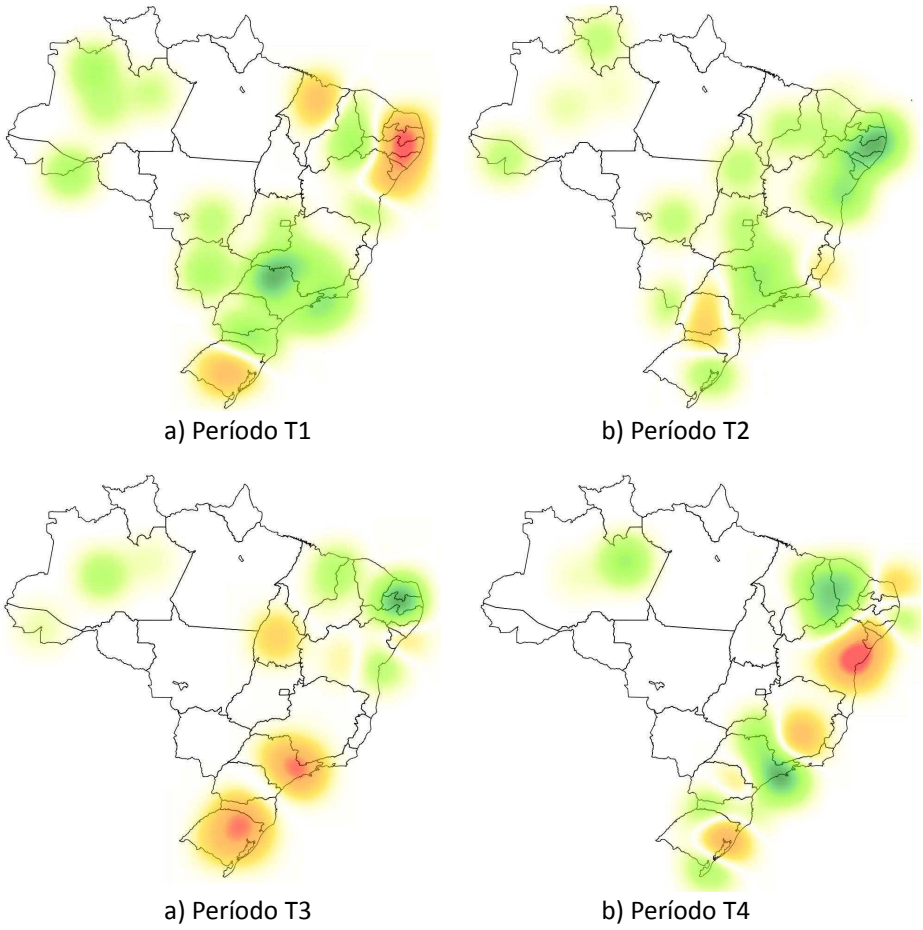


Figura 5.6: Análise do Sentimento Espaço-Temporal: Mapas de Calor dos Períodos Analisados

presso nos tweets referentes a alguns estados do nordeste havia uma orientação semântica negativa e que, durante o jogos, houve mudança na polaridade do sentimento, conforme apresenta o mapa de calor da Figura 5.6 (b). Analisando exclusivamente o gráfico da Figura 5.5, percebe-se que, durante a competição ($T2$), a orientação semântica do sentimento foi predominantemente positiva, e este sentimento positivo foi expresso por quase todas as localizações geográficas. Já no período pós competição ($T3$), percebe-se divergências em relação a polaridade do sentimento em vários estados, onde a maioria dos estados do nordeste apresentou uma orientação semântica do sentimento positiva e nos estados do sul e sudeste apresentaram uma orientação semântica negativa.

5.5 Considerações do Capítulo

Este capítulo apresentou o estudo de caso realizado sobre a Copa das Confederações de 2013 no Brasil, utilizando a abordagem de análise de sentimentos desta pesquisa e teve, como principal objetivo, validar as técnicas utilizadas na abordagem. Os resultados obtidos mostraram que a sumarização espaço-temporal oferece suporte ao tomador de decisões, ao utilizar-se das abordagens de sumarização do sentimento propostas nesta dissertação. Além do mais, os resultados obtidos referentes às técnicas de detecção de polaridade do sentimento foram satisfatórios, considerando o estudo realizado em microtextos escritos no idioma português.

No próximo capítulo, serão apresentadas as considerações finais sobre o trabalho desenvolvido nesta pesquisa, suas contribuições e os trabalhos futuros.

Capítulo 6

Conclusões e Trabalhos Futuros

O volume crescente de conteúdo subjetivo proporcionado pela Web 2.0 através das diversas mídias sociais, tem tornado a análise de sentimentos um campo de pesquisa cada vez mais atrativo, principalmente pela possibilidade de oferecer para as organizações a habilidade de monitorar as opiniões das pessoas em tempo real nas mídias sociais, dando suporte nas tomadas de decisões. Neste trabalho, foi proposta uma abordagem de análise de sentimentos espaço-temporal em microtextos para que, através de técnicas de detecção de polaridade de sentimento e técnicas de GIR, sejam oferecidas possibilidades de sumarização e visualização do sentimento, incluindo as dimensões espacial e temporal.

Para detecção da polaridade de sentimentos foram implementados algoritmos com abordagens léxica e com aprendizado de máquina supervisionado. Na abordagem léxica, dois dicionários de sentimentos disponíveis para o idioma português foram analisados no algoritmo desenvolvido, no entanto, assim como em outros trabalhos semelhantes, as abordagens léxicas foram consideradas inadequadas para textos informais, como os encontrados em tweets. Com relação à abordagem de aprendizado de máquina supervisionado, foram implementadas e comparadas duas abordagens de classificação de texto: Classificação Simples e Classificação Dupla. Em ambas as abordagens a ideia foi eliminar uso de POS Tagging. A realização de POS Tagging aplicado em textos informais é um problema ainda em aberto e seu estudo não fez parte do escopo desta pesquisa. A abordagem de Classificação Dupla mostrou-se mais eficiente quanto ao processo de detecção de polaridade. Isto ocorreu devido à utilização de dois classificadores, onde o primeiro classifica os textos em objetivos e subjetivos (opinativos), funcionando como uma espécie de filtro, e o segundo classifica as polaridades

nas classes Positiva ou Negativa.

Nas abordagens de aprendizado de máquina foram comparados os resultados de dois classificadores: Naive-Bayes e SVM. Considerando os tweets coletados referentes a Copa das Confederações da FIFA, realizada no Brasil em 2013, o melhor resultado da classificação do sentimento foi obtido utilizando o modelo de classificação SVM, resultando em uma F-Measure de 0,873 e precisão de 80%. Estes valores representam bons resultados para os tweets em Português, especialmente se considerar que a polaridade do conteúdo subjetivo nem sempre é consensual. Outros estudos de análise de sentimento aplicados ao idioma inglês, considerando os melhores cenários, atingem uma precisão de cerca de 95% para detecção da polaridade do sentimento.

Além da detecção da polaridade do sentimento, esta pesquisa requereu a utilização de técnicas de Recuperação da Informação Geográfica (GIR) para inferir localizações geográficas através das evidências textuais contidas nos microtextos. Neste sentido, esta pesquisa avaliou a utilização de módulos do GeoSEn para realizar os processos de geoparser (detecção de lugares) e geocoding (avaliação de relevância para definição do escopo geográfico) nos microtextos. Os resultados da avaliação indicaram uma precisão de 92,3% e uma revocação de 47,45%. Do resultado da revocação, têm-se a necessidade de aumentar o escopo geográfico do GeoSEn para considerar níveis abaixo de municípios, como Pontos de Interesse (POI), bairro e ruas, melhorando assim o mecanismo de definição do escopo geográfico.

Nesta abordagem de análise de sentimento, foram utilizadas três formas de sumarização do sentimento: Análise Temporal do Sentimento, Word Clouds do Sentimento e Visualização Espacial do Sentimento. Na Análise Temporal do Sentimento, o objetivo foi gerar gráficos que possibilitem traçar o sentimento ao longo do período escolhido, através dos quantitativos de microtextos com sentimentos positivos e negativos. Estes gráficos permitem compreender a orientação semântica do sentimento geral expresso pela população. Neste sentido, uma ferramenta interativa foi desenvolvida permitindo que os usuários visualizassem os sentimentos em um período de tempo especificado. Ainda no processo de sumarizar o sentimento detectado, este trabalho abordou a geração de Word Clouds (Nuvens de palavras) para auxiliar na compressão do sentimento detectado em um certo período ou data especificada. Foi proposto ainda que a geração da Word Cloud fosse realizada em datas cujas quantidades de mensagens opinativas foram atípicas (outliers) para o período observado, visto que a word

cloud pode auxiliar na visualização dos termos mais frequentemente citados, permitindo a compreensão dos fatos. Por fim, esta pesquisa abordou a geração de mapas geográficos do sentimento, nos quais o quantitativo dos sentimentos detectados nas localizações geográficas são gerados através dos indicadores de intensidades no mapa de calor. Com esta abordagem, é possível visualizar, graficamente em mapas, as mudanças de opiniões ocorridas nas localizações geográficas ao longo do tempo.

6.1 Contribuições

As principais contribuições desta pesquisa são:

- Implementação de técnicas de classificação de sentimentos, que quando aplicadas em microtextos escritos no idioma português, apresentaram resultados de acurácia, precisão e revocação considerados satisfatórios;
- Comparação de abordagens de técnicas de análise de sentimentos utilizando-se de microtextos escritos no idioma português;
- Avaliação de técnicas de GIR para inferir referências geográficas em microtextos;
- Proposta de sumarização dos sentimentos explorando a dimensão espacial e temporal;
  - A utilização de mapas de calor para sumarizar o sentimento é uma alternativa de sumarização que ainda não foi explorada na literatura;

6.2 Trabalhos Futuros

Através dos resultados obtidos nesta pesquisa, observaram-se os seguintes trabalhos futuros:

- Implementar uma técnica de visualização espacial para destacar as mudanças de opiniões ocorridas sobre o espaço e tempo, incluindo os usuários dos tweets h_{t_i};
- Explorar outras entidades dos tweets além da mensagem textual, tanto para prover melhorias na acurácia da polaridade do sentimento quanto na detecção de regiões geográficas. A exploração pode ser feita, por exemplo, utilizando-se informações dos

seguidores de usuários ou do histórico de outros tweets enviados, bem como a utilização das informações geográficas contidas em alguns tweets;

- Explorar séries temporais para prever os sentimentos nas diversas regiões geográficas;
- Utilizar uma abordagem de análise de sentimentos em nível de aspectos ($a_j^{t_i}$), para através do Reconhecimento de Entidades Nomeadas (NER) possibilitar a detecção da polaridade no nível das características da entidade analisada;
- Aprimorar as heurísticas de georeferenciamento do GeoSEn para melhorar os índices de revocação.

Bibliografia

[1] Mikalai Tsytsarau and Themis Palpanas. Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, 24(3):478–514, May 2012.

[2] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[3] Jianwei Zhang, Yukiko Kawai, Tadahiko Kumamoto, and Katsumi Tanaka. A novel visualization method for distinction of web news sentiment. In Gottfried Vossen, DarrellD.E. Long, and JeffreyXu Yu, editors, *Web Information Systems Engineering - WISE 2009*, volume 5802 of *Lecture Notes in Computer Science*, pages 181–194. Springer Berlin Heidelberg, 2009.

[4] Eivind Bjø rkelund, Thomas H Burnett, and Kjetil Nø rvåg. A study of opinion mining and visualization of hotel reviews. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services - IIWAS '12*, page 229, New York, New York, USA, 2012. ACM Press.

[5] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82, April 2013.

[6] Cláudio Elízio Calazans Campelo. Geosen: um motor de busca com enfoque geográfico. Master's thesis, Universidade Federal de Campina Grande, 2008.

[7] Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10*, European Language Resources Association (ELRA), pages 1320–1326, 2010.

[8] Duarte Choon Dias. Text Mining Methods for Mapping Opinions from Georeferenced Documents. Master's thesis, Universidade Técnica de Lisboa, 2012.

[9] M. G. de Oliveira, C. S. Baptista, C. E. C. Campelo, J. A. M. Acioli Filho, and A. G. R. Falcão. Automated production of volunteered geographic information from social media. In *XV Brazilian Symposium on Geoinformatics (in press)*, Campos do Jordão, SP, 2014. SBC.

[10] Andreas Auinger and Martin Fischer. Mining consumers and opinions on the web. *FH Science Day*, pages 410–419, 2008.

[11] Tim O'Reilly. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, 1(65):17–37, 2007.

[12] Magdalini Eirinaki, Shamita Pisal, and Japinder Singh. Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, 78(4):1175–1184, July 2012.

[13] Daniel Appelquist, Dan Brickley, Melvin Carvahlo, Renato Iannella, Alexandre Passant, and Christine Perey. A Standards-based, Open and Privacy-aware Social Web, 2010.

[14] Jeff Zabin and Alex Jefferies. Social media monitoring and analysis: Generating consumer insights from online conversation. *Aberdeen Group Benchmark Report*, 2008.

[15] Bing Liu. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing*, A Chapman amp Hall book, chapter 28, pages 1–38. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN, 2010.

[16] Bing Liu. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, May 2012.

[17] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and TrendsÂ® in Information Retrieval*, 2(2):1–135, 2008.

[18] Moshe Koppel and Itai Shtrimberg. Good News or Bad News ? Let the Market Decide. *AAAI Spring Symposium on Exploring Attitude and Affect in Tex*, pages 86–88, 2004.

[19] Neil O'Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin, and Alan F. Smeaton. Topic-dependent sentiment analysis of financial blogs. *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion - TSA '09*, page 9, 2009.

[20] Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. PolariCQ. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, page 1945, New York, New York, USA, 2012. ACM Press.

[21] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting Elections with Twitter : What 140 Characters Reveal about Political Sentiment. *Proceedings of the fourth international aaai conference on weblogs and social media*, pages 178–185, 2010.

[22] Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM'12*, page 63, New York, New York, USA, 2012. ACM Press.

[23] Tun Thura Thet, Jin-Cheon Na, Christopher S.G. Khoo, and Subbaraj Shakthikumar. Sentiment analysis of movie reviews on discussion boards using a linguistic approach. *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion - TSA '09*, page 81, 2009.

[24] Thomas Hoberg Burnett and Eivind Bjø rkelund. Temporal Opinion Mining. Technical Report June, Norwegian University of Science and Technology, Trondheim, 2012.

[25] Anuj Sharma and Shubhamoy Dey. A boosted svm based sentiment analysis approach for online opinionated text. In *Proceedings of the 2013 Research in Adaptive and Convergent Systems*, RACS '13, pages 28–34, New York, NY, USA, 2013. ACM.

[26] Jerry R. Hobbs and Ellen Riloff. Information extraction, in handbook of natural language processing. *Chapman & Hall CRC Press*, 2010.

[27] Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computati-*

onal Linguistics: Human Language Technologies - Volume 1, HLT '11, pages 359–367, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[28] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 129–136, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[29] Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[30] Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'05, pages 486–497, Berlin, Heidelberg, 2005. Springer-Verlag.

[31] Anuj Sharma and Shubhamoy Dey. A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium on - RACS '12*, page 1, New York, New York, USA, 2012. ACM Press.

[32] Diego Tumitan Karin Becker. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. In *Tutorials no 28º Simpósio Brasileiro de Banco de Dados*, 2013.

[33] Alec Go, Lei Huang, and Richa Bhayani. Twitter Sentiment Analysis. *CS224N - Final Project Report for Spring 2008/2009*, 2009.

[34] Ludmila I. Kuncheva. On the optimality of naive bayes with dependent binary features. *Pattern Recogn. Lett.*, 27(7):830–837, 2006.

[35] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[36] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 2000.

[37] Øyvind Vestavik. Geographic Information Retrieval: An Overview. Dept. Computer and Information Science, Norwegian University of Technology and Science, Trondheim, Norway, 2008.

[38] Alexander Markowetz, Yen yu Chen, and Torsten Suel. Design and implementation of a geographic search engine. In *In 8th Int. Workshop on the Web and Databases (WebDB*, 2005.

[39] Jochen L. Leidner and Michael D. Lieberman. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11, July 2011.

[40] A. M.; Medeiros J. S. Câmara, G.; Monteiro. Fundamentos epistemológicos da ciência da geoinformação. Introdução ao geoprocessamento, INPE, 2000.

[41] Karla A. V. Borges. Frederico Fonseca, Max Egenhofer. Ontologias e interoperabilidade semântica entre sigs. In *GeoInfo 2011*, 2011.

[42] Cleber Gouvêa. Uma Abordagem para o Enriquecimento de Gazetteers a partir de Notícias visando o Georreferenciamento de Textos na Web. Master's thesis, Universidade Católica de Pelotas, 2009.

[43] David M. W. Powers. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia, 2007.

[44] Pedro Henrique Calais Guerra, Adriano Veloso, Wagner Meira, Jr., and Virgílio Almeida. From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 150–158, New York, NY, USA, 2011. ACM.

[45] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, CIKM '11, page 1031, New York, New York, USA, 2011. ACM Press.

[46] Marlo Souza and Renata Vieira. Sentiment Analysis on Twitter Data for Portuguese Language. *PROPOR'12 Proceedings of the 10th international conference on Computational Processing of the Portuguese Language*, pages 241–247, 2012.

[47] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[48] Yung-Ming Li and Tsung-Ying Li. Deriving Marketing Intelligence over Microblogs. In *2011 44th Hawaii International Conference on System Sciences*, pages 1–10. IEEE, January 2011.

[49] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks - COSN '13*, pages 27–38, New York, 2013. ACM Press.

[50] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, volume 10, pages 79–86, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[51] Marcirio Chaves, Larissa De Freitas, Marlo Souza, and Renata Vieira. PIRPO: An Algorithm to Deal with Polarity in Portuguese Online Reviews from the Accommodation Sector. *Natural Language Processing and Information Systems*, 7337:1–5, 2012.

[52] Luís Sarmento, Paula Carvalho, Mário J. Silva, and Eugénio de Oliveira. Automatic creation of a reference corpus for political opinion mining in user-generated content.

In *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion - TSA '09*, page 29, New York, 2009. ACM Press.

[53] Diego Tumitan and Karin Becker. Tracking Sentiment Evolution on User-Generated Content : A Case Study on the Brazilian Political Scene. *Simpósio Brasileiro de Banco de Dados - SBBD 2013*, pages 1–6, 2013.

[54] M. Souza, R. Vieira, R. Busetti, D. andChishman, and I. M Alves. Construction of a portuguese opinion lexicon from multiple resources. In *8th Brazilian Symposium in Information and Human Language Technology*, 2012.

[55] Mário J. Silva, Paula Carvalho, and Luís Sarmento. Building a sentiment lexicon for social judgement mining. In Helena Caseli, Aline Villavicencio, Antônio Teixeira, and Fernando PerdigÃ£o, editors, *Computational Processing of the Portuguese Language*, volume 7243 of *Lecture Notes in Computer Science*, pages 218–228. Springer Berlin Heidelberg, 2012.

[56] Paula Nascimento, Rodrigo Aguas, Débora De Lima, Xiao Kong, and Bruno Osiek. Análise de sentimento de tweets com foco em notícias. In *I Brazilian Workshop on Social Network Analysis and Mining*, 2012.

[57] Janaína Gomide, Adriano Veloso, Wagner Meira, Jr., Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 3:1–3:8, New York, NY, USA, 2011. ACM.

[58] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 06*, pages 417–422, 2006.

[59] G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6), 2012.

[60] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and*

Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[61] John C. Platt. Fast training of support vector machines using sequential minimal optimization. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in kernel methods*, chapter Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[62] Cláudio Elízio Calazans Campelo and Cláudio de Souza Baptista. A model for geographic knowledge extraction on web documents. In CarlosAlberto Heuser and Günther Pernul, editors, *Advances in Conceptual Modeling - Challenging Perspectives*, volume 5833 of *Lecture Notes in Computer Science*, pages 317–326. Springer Berlin Heidelberg, 2009.

[63] Natalia V. Andrienko, Gennady L. Andrienko, and Peter Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *J. Vis. Lang. Comput.*, 14(6):503–541, 2003.