UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

THYAGO ROMAGNA BENDO

ANÁLISE ESPAÇO-TEMPORAL DE DADOS GEORREFERENCIADOS DE CASOS DE DENGUE IDENTIFICADOS A PARTIR DE REDE SOCIAL NO ESTADO DO PARANÁ

TRABALHO DE CONCLUSÃO DE CURSO

MEDIANEIRA

2019

THYAGO ROMAGNA BENDO

ANÁLISE ESPAÇO-TEMPORAL DE DADOS GEORREFERENCIADOS DE CASOS DE DENGUE IDENTIFICADOS A PARTIR DE REDE SOCIAL NO ESTADO DO PARANÁ

Trabalho de Conclusão de Curso apresentado ao Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de "Bacharel em Computação".

Orientador: Prof. Dr. Alan Gavioli

Co-orientador: Prof. Dr. Claudio Leones Bazzi

MEDIANEIRA

2019

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Diretoria de Graduação e Educação Profissional
Coordenação do Curso de Ciência da Computação

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

TERMO DE APROVAÇÃO

ANÁLISE ESPAÇO-TEMPORAL DE DADOS GEORREFERENCIADOS DE CASOS DE DENGUE IDENTIFICADOS A PARTIR DE REDE SOCIAL NO ESTADO DO PARANÁ

Por

THYAGO ROMAGNA BENDO

Este Trabalho de Conclusão de Curso foi apresentado às 11:10h do dia 1 de julho de 2019 como requisito parcial para a obtenção do título de Bacharel no Curso de Ciência da Computação, da Universidade Tecnológica Federal do Paraná, Câmpus Medianeira. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. Alan Gavioli
UTFPR - Câmpus Medianeira

Prof. Arnaldo Candido Junior
UTFPR - Câmpus Medianeira

Prof. Evando Carlos Pessini
UTFPR - Câmpus Medianeira

Prof. Claudio Leones Bazzi
UTFPR - Câmpus Medianeira

A folha de aprovação assinada encontra-se na Coordenação do Curso.

RESUMO

BENDO, Thyago Romagna. ANÁLISE ESPAÇO-TEMPORAL DE DADOS GEORREFERENCIADOS DE CASOS DE DENGUE IDENTIFICADOS A PARTIR DE REDE SOCIAL NO ESTADO DO PARANÁ. 72 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2019.

O crescente uso de redes sociais pela população brasileira permite que uma quantidade cada vez maior de dados seja gerada. Esses dados possibilitam que diversos temas possam ser objetos de análise. Este trabalho objetivou analisar dados obtidos de uma rede social (Twitter) buscando identificar conteúdos que indiquem casos de dengue, por meio do uso de um modelo classificador desenvolvido após testes de opções de amostragem e pré-processamento e de algoritmos J48, Máquinas de Vetor de Suporte, Máxima Entropia e Bayesiano Ingenuo Multinomial. Os casos identificados foram utilizados em uma nova etapa de análise, onde a partir da localização geográfica de cada *tweet* georreferenciado foram definidas as suas respectivas regiões por meio do agrupamento por densidade do algoritmo BDSCAN, formando assim o conjunto de cidades consideradas de risco pela identificação por rede social. Para verificar se os dados obtidos pelas análises foram precisos, os resultados foram comparados com os dados de casos de dengue registrados pelo Ministério da Saúde, com as comparações contando com diferentes granularidades espaciais e temporais, de forma a distinguir onde se encontrava a aplicação de maior eficácia. Os resultados permitiram determinar um modelo de classificar *tweets* que indicam casos de dengue com 91% de acerto e determinar o padrão da eficiência da identificação se relacionando diretamente com a quantidade de dados disponíveis para análise, podendo ser a abordagem espaço-temporal aplicável de acordo com o interesse de precisão mínima, variando entre 23% e 80% de acordo com a granularidade temporal.

Palavras-chave: rede social, dengue, mineração de dados

ABSTRACT

BENDO, Thyago Romagna. QUADROTOR. 72 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2019.

The rising use of social media by the Brazilian population allows generating an increasing quantity of data. This data allows analyzing many subjects as a study case. This project aims to analyze de data from social media (Twitter) searching to identify contents which indicate dengue cases, by the use of a classifier model developed with sampling and preprocessing tests and analysis based on J48, Support Vector Machine, Maximum Entropy and Multinomial Naive Bayes. A new step used this social media identified cases, using their geographical localization in a DBSCAN density cluster to define the regions with more dengue incidence, forming the social media identified risk regions. To verify the results precision, the results were compared with the Health Ministry official data, making comparisons in different spatial-temporal granularities, to discern where the identification efficacy was. The results allowed to determinate the dengue cases indicative tweets classifier model with 91% of precision. They determinate too the pattern of the efficiency, relating directly with the available data volume, with the spatial-temporal analysis being able to be applied according to the minimum precision desired, varying between 23% and 80% according to the spatial granularity.

Keywords: social media, dengue, data mining

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ACID	Atomicidade, Consistência, Isolamento e Durabilidade
API	Interface de Programação de Aplicações
JSON	Notação de Objetos JavaScript
KDD	Knowledge Discovery from Data
LSA	Latent Semantic Analysis
NLP	Processamento de Língua Natural
SIG	Sistema de Informação Geográfica
SGBD	Sistema Gerenciador de Banco de Dados
SQL	Structured Query Language
SVM	Máquinas de Vetores de Suporte
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

A dengue é uma doença transmitida por meio do mosquito Aedes Aegypti, que está presente em muitos países, infectando globalmente entre 50 e 200 milhões de pessoas por ano. As Américas foram responsáveis por 14% desses casos, com mais da metade desses casos atribuídos a Brasil e México (BHATT et al., 2013). No Brasil os dados são fornecidos pelo Ministério da Saúde em boletins epidemiológicos, e segundo os dados das publicações de Brasil (2015), Brasil (2016), Brasil (2017), foram registrados 3.421.367 casos no triênio, sendo a doença epidemiológica que mais infecta pessoas no país.

Como abordado por Fernando e Jesus (2010), o estudo epidemiológico está estreitamente ligado a abordagem espacial. Essa abordagem pode ser aprimorada pelo uso de Sistema de Informação Geográfica (SIG), que permite uma abordagem computacional do estudo do espaço. Um SIG é um sistema que pode ser entendido como algo único, já que por tratar sobre informações geográficas, diferentemente dos sistemas prévios a ele, pode ser aplicado em muitas áreas, como informações demográficas e de trânsito (TIAN, 2016).

Para essa abordagem computacional é necessária a existência de dados em uma quantidade significativa, para que os dados tenham uma proximidade maior da realidade. Para isso é necessário que esses dados sejam obtidos a partir de algo com grande disseminação na população. Dois itens com grande presença entre a população brasileira são: *smartphones* e computadores. A distribuição desses itens de forma *per capita* são respectivamente de 106% e 83%, segundo dados da pesquisa MEIRELLES (2018). Essa grande presença de equipamentos com acesso à Internet permite o grande acesso a redes sociais, que são ambientes onde as pessoas compartilham os acontecimentos de sua vida diariamente.

A partir da disponibilização de dados da rede social e por meio de técnicas de extração de informação, é possível obter dados relevantes a partir do conteúdo gerado a partir das informações dos usuários (YOO et al., 2018). Essas técnicas de extração de informação podem envolver desde análises sobre o texto até mesmo imagens, podendo ser desde uma classificação do conteúdo até uma análise do sentimento envolvido no contexto, a qual pode identificar os sentimentos do autor do conteúdo com o objeto presente no mesmo.

Como já demonstrado em trabalhos similares por Gomide et al. (2011) e Sousa et

al. (2018), é possível obter informações sobre a presença de dengue de forma espacial, e essas informações já mostraram um grau de relação com os casos de dengue identificados oficialmente.

1.1 OBJETIVO GERAL

O objetivo deste trabalho é analisar a aplicabilidade da identificação geográfica de casos de dengue em todo o estado do Paraná por meio do Twitter, comparando dados reais de todas as cidades do estado com dados obtidos por meio de mineração de sentimentos.

1.2 OBJETIVOS ESPECÍFICOS

- Elaborar uma base de dados que armazene registros de casos de dengue de todas as cidades do Paraná com as respectivas informações geográficas da cidade de ocorrência;
- Desenvolver um classificador que seja capaz de identificar relações pessoais com a dengue em dados de rede social, obtidos por meio da Interface de Programação de Aplicações (API) de dados do Twitter, utilizando a classificação que considere o sentimento no texto em relação a palavra dengue;
- Estabelecer a base de dados geográfica dos casos de dengue identificados a partir das informações obtidas pela classificação de *tweets* com a palavra dengue, restritos a região do estado do Paraná;
- Realizar uma análise comparativa em termos de localização geográfica entre os dados reais de casos de dengue e os dados obtidos por meio da mineração de rede social, para verificar possíveis correspondências.

1.3 JUSTIFICATIVA

Sendo a dengue uma doença que no triênio entre 2015 e 2017 causou 32.629 casos com sinais de alarme e 1.814 óbitos segundo dados dos boletins epidemiológicos Brasil (2015), Brasil (2016), Brasil (2017), verifica-se a necessidade de entender a presença da doença no país.

Como primeira alternativa para isso existem os dados oficiais, entretanto Codeco et al. (2016) mostram que existe um atraso de 2 semanas na mensuração da situação epidemiológica da dengue, que precisa ser corrigida através do uso de uma equação, que apresenta erros na previsão.

A partir da análise de Kwak et al. (2010) de que uma rede social, nesse caso especificamente do Twitter, pode funcionar como uma fonte relevante de notícias, esse meio pode se mostrar um caminho viável para mensurar a situação da dengue.

Dessa forma se mostra que é relevante analisar a aplicabilidade da abordagem de identificação de casos de dengue, com a determinação das cidades, por meio de dados de rede social, obtendo uma alternativa que pode ser aplicada em um tempo menor do que os dados oficiais, para que possa atuar como um complemento as análises de dados reais já existentes.

1.4 ORGANIZAÇÃO DO DOCUMENTO

Esse documento será organizado da seguinte forma. O Capítulo 2 irá apresentar o referencial teórico, a metodologia utilizada se encontra no Capítulo 3, nele sendo descritas todas as etapas para o desenvolvimento do projeto. No Capítulo 4 são apresentados os resultados obtidos, tendo como sequencia o Capítulo 5 onde serão discutidos os resultados do trabalho e, por fim, no Capítulo 6 encontra-se a conclusão da proposta do trabalho de conclusão de curso.

2 REFERENCIAL TEÓRICO

Esse capitulo irá abordar os assuntos referentes a todo o referencial teórico, se dividindo em seções para o Georreferenciamento 2.1; Mineração de Dados 2.2; Mineração de Dados Georreferenciados 2.3; Análise de Sentimentos 2.4 e Trabalhos Correlatos 2.5

2.1 GEORREFERENCIAMENTO

A cartografia é um termo oriundo do século XIX, entretanto o ato de se trabalhar com mapas já é praticado pela humanidade há muito mais tempo, e foi muito importante em algumas etapas do desenvolvimento humano (CRAMPTON, 2011).

Entretanto, a geografia deixou de ser somente sobre os mapas e passou a se relacionar com a percepção de como é o mundo, o que é feito por meio de observação, medição e análise dos dados oriundos do ambiente, para um melhor desenvolvimento dessa atividade foi utilizado a computação, o que deu origem aos Sistemas de Informação Geográfica (DALE, 2014).

2.1.1 Sistema de Informação Geográfica

Um SIG é utilizado em diversas aplicações que necessitam de informações espaciais, sendo importante desde a coleta e o armazenamento de dados até as etapas de análise e visualização desses dados, fornecendo informações precisas sobre posicionamentos na Terra, o que é muito útil em localizações de amostras, posicionamento de parcelas de terra e em linhas comerciais (HEYWOOD et al., 2011).

Já quanto as características dos dados de um SIG, O'sullivan e Unwin (2014) as

dividem em duas categorias, os dados vetoriais onde são armazenadas as coordenadas dos pontos, que podem ainda formar linhas e polígonos e dados *raster*, que são relacionados com pequenas unidades em *grid*, chamadas píxeis, e que são formadas a partir de imagens da superfície da Terra. Enquanto o primeiro é mais utilizado para trabalhos com mapas o segundo é mais utilizado no sensoriamento remoto (BIVAND et al., 2008).

Na Figura 1 é representado um exemplo de dado vetorial, no caso um mapa do Estado do Paraná, com a cidade de Medianeira como destaque. Já na Figura 2 o exemplo é de dados *raster*, no caso uma imagem do rio Paraná na divisa entre Brasil e Paraguai.

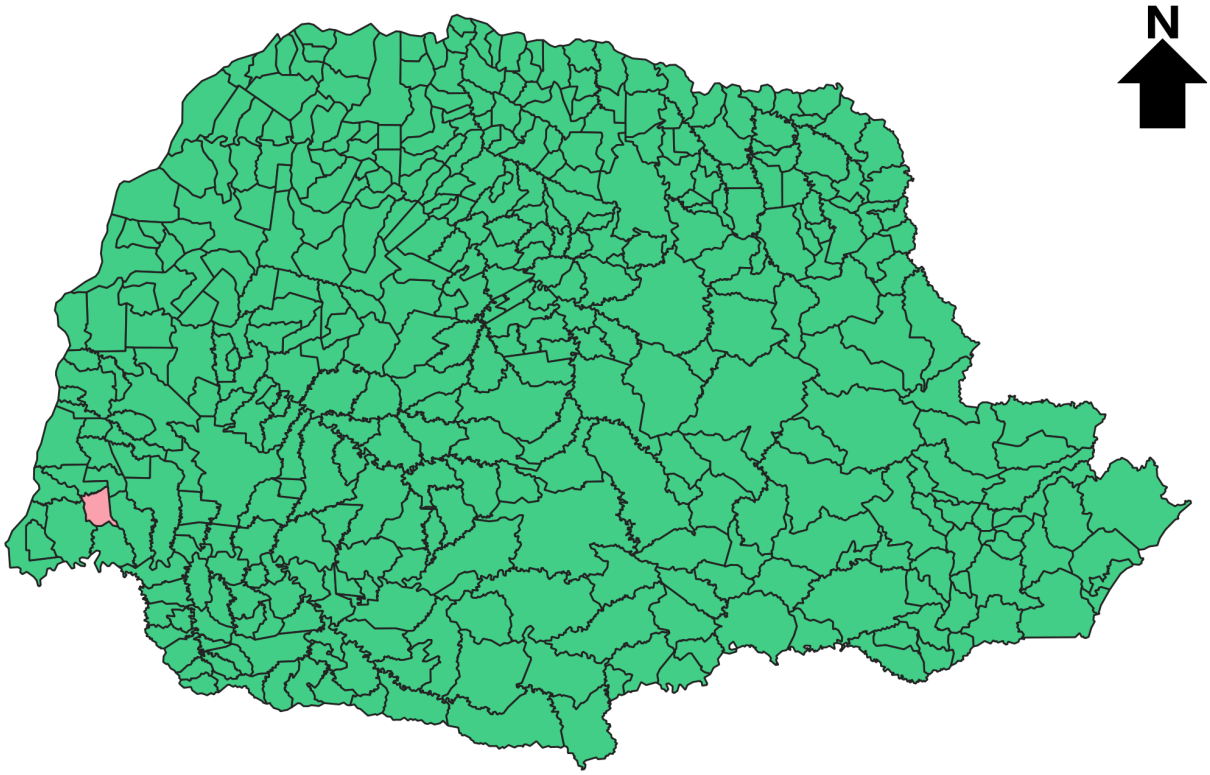

Figura 1 – Exemplo de dado vetorial.
Fonte: Autoria própria

Para determinar as posições geográficas no Globo Terrestre existem duas abordagens que podem ser utilizadas, as coordenadas esféricas que fornecem uma representação em três dimensões e coordenadas projetadas que representam a Terra em duas dimensões com uma série de cálculos (TIAN, 2016). Na Figura 3 é apresentado um mapa construído em duas dimensões e com as indicações para coordenadas esféricas.

Entretanto ambas as coordenadas enfrentam problemas comuns na representação geográfica, pois os objetos normalmente são multidimensionais, nem sempre com geometrias simples, podendo ser inclusive um fractal, além de poderem ter bordas confusas e/ou indeterminadas, o que dificulta muito a representação. Ainda há fatores como a escala, que

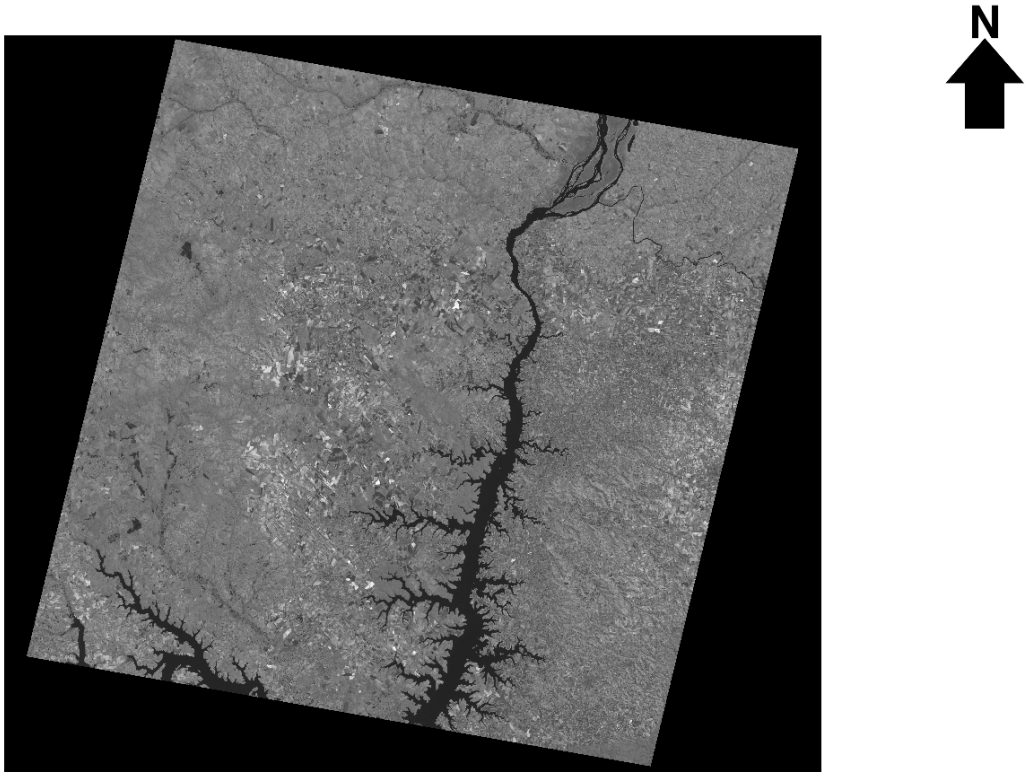

Figura 2 – Exemplo de dado raster.

Fonte: Autoria própria

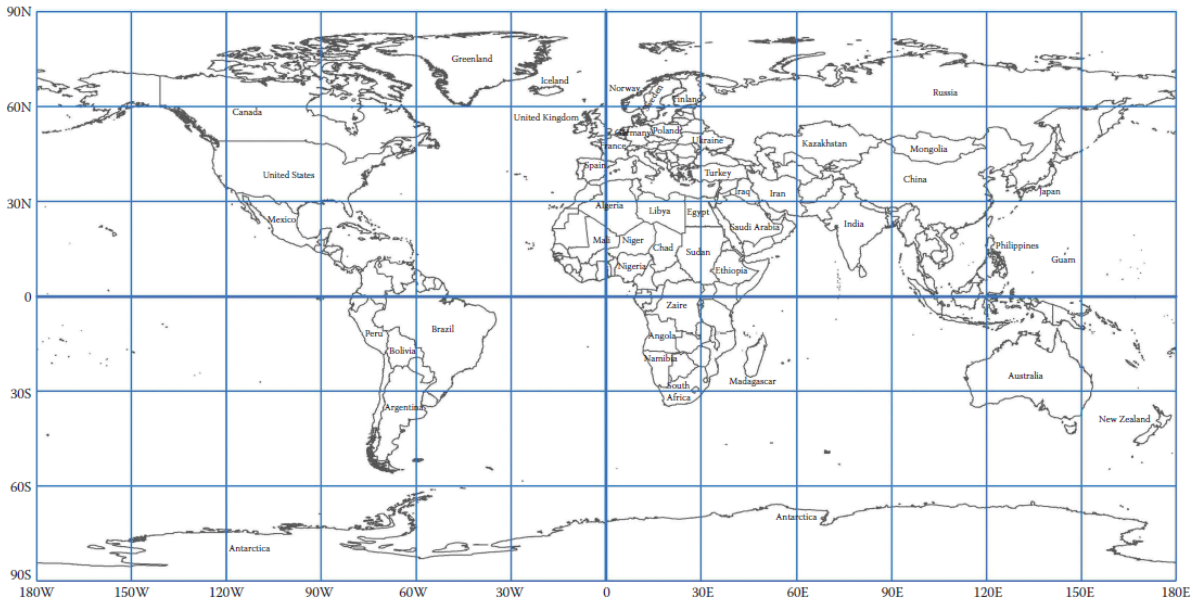

Figura 3 – Exemplo de coordenadas.

Fonte: (TIAN, 2016)

pode interferir muito na representação, além das diferentes formas de se mensurar as distâncias (O'SULLIVAN; UNWIN, 2014).

Por isso são utilizados diferentes meios de ser representar as coordenadas, dependendo da aplicação desejada, esses podem preservar ângulos, áreas ou equidistância, a escolha por um dos itens tende a prejudicar a representação dos outros, por isso é possível utilizar ainda uma representação ajustada, que não é a ideal em nenhum dos itens, mas possui um equilíbrio entre diferentes itens (LONGLEY et al., 2005).

Para a representação computacional dessas coordenadas são utilizados modelos matemáticos *datums*, os modelos mais comuns são: *World Geodetic Survey* de 1984; Elipsoide de Clarke de 1866; *Global Reference System* de 1980 e *International Terrestrial Reference Frame* de 1997. Esses diferentes modelos são adaptados para a obtenção de uma melhor localidade em cada região (TIAN, 2016).

Esses modelos são utilizados para a visualização, armazenamento e cálculos que considerem a posição geográfica, sendo que em muitos casos o último item é implementado juntamente ao armazenamento, como melhor será explorado a seguir.

2.1.2 Base de Dados Geográficos

O armazenamento é um tema-chave quando se trata dos dados geográficos, pois é importante que toda a informação obtida seja preservada e esteja disponível para acesso, e para isso existem diferentes abordagens como o uso da estrutura de banco de dados relacionais, de uma base de dados georreferenciada ou de uma base de dados orientada a objetos (DIXON et al., 2015).

Como vantagens para o armazenamento que utiliza a estrutura de um banco de dados relacional está todo o legado que o banco fornece, como a Atomicidade, Consistência, Isolamento e Durabilidade (ACID). Nesse caso geralmente armazenam os dados geográficos em uma tabela ou coluna específica (HEYWOOD et al., 2011).

Para o armazenamento em uma base de dados georreferenciada, normalmente, são utilizadas as características de uma das outras bases para a integridade dos dados, e normalmente a sua grande vantagem é ser um ambiente mais completo, permitindo todos os tipos de dados já descritos com diversas formas de inserção e acesso, para ser acessível para diferentes sistemas (DIXON et al., 2015).

Já uma base de dados orientada a objetos fornece as características dos modelos de orientação a objetos, de forma que cada item é tratado como um objeto e a questão geográfica como mais um atributo do objeto que possui essa característica (CÂMARA et al., 1996).

Muitas dessas bases ainda fornecem elementos para facilitar o trabalho com elementos geográficos, como por exemplo o cálculo de uma distância considerando a curvatura terrestre, que exigiria um conhecimento muito maior do utilizador, mas em muitos casos pode ser executada com a chamada de uma função (CORTI et al., 2014), como no exemplo da extensão espacial PostGIS do banco de dados relacional PostgreSQL:

```
float ST_Distance(geometry g1, geometry g2);
```

Esse tipo de função pode ser utilizado em vários casos, que vão desde uma nova função, uma *trigger* ou mesmo em uma consulta simples, como é o caso visto abaixo, onde é calculada a distância entre um ponto e uma linha:

```
SELECT ST_Distance(
'SRID=4326;POINT(-72.01 42.35)'::geometry,
'SRID=4326;LINESTRING(-72.12 42.45, -72.12 42.15)'::geometry
);
```

2.1.3 Aplicações de Sistemas de Informação Geográfica

Os SIGs podem ser aplicados em diversas áreas do conhecimento e para diversos fins, entre alguns desses fins pode-se destacar aplicações no campo da saúde, como feito nos trabalhos de Guagliardo (2004), Ghosh e Guha (2013), Allen et al. (2016), no campo da agricultura, como nos trabalhos de Zhang et al. (2002), Giri et al. (2011), e em planejamento urbano, como visto nos trabalhos de Xiao et al. (2006), Bathrellos et al. (2017).

Os exemplos citados anteriormente abordam diversos meios de obtenção dos dados, dentre eles estão presentes a obtenção de imagens por meio de satélites, mapas e redes sociais, as quais se tornaram um tópico de interesse devido ao seu crescimento nos últimos anos (MUNINGER et al., 2019).

É comum entre as redes sociais o uso de funções que registrem a localização do usuário no momento em que o mesmo realiza alguma ação dentro do ambiente da rede social, sendo que a forma do uso do georreferenciamento e a característica da interação varia de acordo com a política de funcionamento da rede social (SUI; GOODCHILD, 2011a; TOEPKE, 2016).

Independentemente da maneira como é utilizado o georreferenciamento na rede social o ato sempre é de unir a posição geográfica a algum conteúdo do usuário, unindo assim o georreferenciamento a mais um padrão de informação (CHENG et al., 2010).

Essa característica permite o desenvolvimento de uma nova abordagem do georreferenciamento, pois as redes sociais permitem uma relação de proporções muito diferentes das existentes anteriormente entre a posição geográfica e a mídia produzida (SUI; GOODCHILD, 2011b).

2.2 DESCOBERTA DE CONHECIMENTO

Nesta seção será apresentado o conceito da descoberta de conhecimento, seu funcionamento e suas principais técnicas e exemplos de aplicação.

2.2.1 Introdução

Fayyad et al. (1996) expõe a diferença entre a Descoberta de Conhecimento ou Extração de Conhecimento (KDD) e a mineração de dados, pois embora muitas vezes sejam erroneamente tratadas como sinônimos, a mineração de dados é apenas uma das etapas de KDD, como mostrado na Figura 4 (HAN et al., 2012).

2.2.2 Seleção

A seleção é a primeira etapa do processo, sendo a responsável pela obtenção dos dados diretamente da fonte alvo para a extração do conhecimento. Han et al. (2012) classifica em 3 as possibilidades de origem para os dados, com essas podendo ser base de dados, *data warehouse*, melhor descrito em (KIMBALL; ROSS, 2011), e dados transacionais, sendo que a origem dos dados tende a caracterizar o posterior processo de mineração.

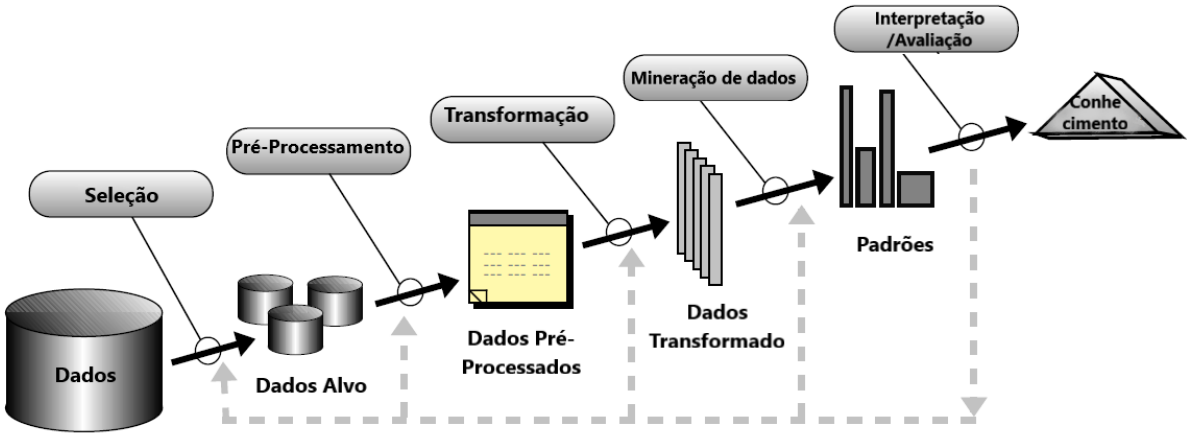

Figura 4 – Onde ocorre a mineração de dados na extração de conhecimento.
Fonte: Adaptado de Fayyad et al. (1996)

Os dados com origem de base de dados adquirem muitas características originárias do seu respectivo banco de dados. No caso de banco de dados relacionais a ACID. Além disso existe o aspecto da forma de obtenção dos dados, que ocorre através do uso de consultas, no caso de bancos de dados relacionais a consulta é feita através de Linguagem de Consulta Estruturada (SQL), o que a permite realizar as operações relacionais, como junções. Essa estrutura fornecida pelo banco de dados permite uma mineração na busca por tendências e padrões (CIOS et al., 2007; HAN et al., 2012).

No caso da origem dos dados ser um *data warehouse*, existe uma facilidade na mineração visando atributos chave, já que esses são os atributos preferíveis na criação de um *data warehouse*, além de que sua estrutura em cubo, com múltiplas dimensões e granularidades, com os dados podendo ser uma agregação de diversas fontes, permite que o uso da mineração na descoberta de padrões de uma forma mais ampla, sendo de grande uso para facilitar tomadas de decisões (CIOS et al., 2007; HAN et al., 2012; FAYYAD, 1997).

Já os dados transacionais permitem análises sobre a forma que algo ocorre, e tem uma diversidade de origem maior, pois os dados podem ser de transações de venda, bom como de cliques do usuário em uma página na Internet. Essa é uma grande fonte de dados para reconhecer padrões entre as ações do usuário, como compra conjuntas ou onde o usuário tende a clicar mais, na Figura 5 é possível visualizar um exemplo de dados transacionais, onde cada transação possui um número identificador e os itens presentes na mesma, que são representados pelas letras de ’a’ à ’e’, podendo representar (BRAMER, 2016; HAN et al., 2012).

Ainda existem diversas outras possíveis origens de dados para a mineração de dados,

Número da Transação	Itens da Transação
1	{a, b, c}
2	{a, b, c, d, e}
3	{b}
4	{c, d, e}
5	{c}
6	{b, c, d}
7	{c, d, e}
8	{c, e}

Figura 5 – Exemplo de dados transacionais.

Fonte: Adaptado de Bramer (2016)

como dados transmitidos em tempo real, geográficos, de engenharia, de hipertexto e multimídia, de grafos e rede, da Internet e de muitas outras fontes, cada um com uma forma de estrutura diferente e com diferentes possibilidades para a mineração e descoberta de conhecimento (HAN et al., 2012).

2.2.3 Pré-processamento

Na sequência tem-se o início da etapa de pré-processamento, que segundo Aggarwal (2015) pode ser considerada a etapa mais crucial de todo o processo de KDD, pois muitas vezes os dados estão incompletos, incorretos ou com um ruído muito alto e inconsistentes, o que atrapalhará o resto do processo de KDD. Assim o pré-processamento tem seu início logo após a obtenção dos dados e ocorre por meio dos processos de limpeza, integração e redução dos dados.

O processo de limpeza dos dados consiste em preencher atributos vazios, reduzir o ruído nos dados, identificar e remover *outliers* e resolver inconsistências. Todas essas atividades

podem ocorrer através de diferentes técnicas, sendo importante identificar a técnica que melhor se adapta ao problema.

Já o processo de integração dos dados se faz necessário quando se deseja utilizar múltiplas fontes de dados para a descoberta de conhecimento, e entre as dificuldades desse processo estão além da possibilidade de integrar diferentes tipos de origem de dados, como bases de dados, cubos e arquivos, tem-se a dificuldade de atributos poderem ter nomes diferentes e terem valores diferentes, devido a abreviações e erros de digitação.

Por sua vez o processo de redução dos dados consiste em reduzir o volume dos dados obtendo como resultado quase o mesmo valor para a análise posterior, de forma a tornar os processos posteriores menos custosos computacionalmente. Esse processo pode ocorrer tanto por meio da redução da dimensionalidade quanto por meio da redução numérica (HAN et al., 2012).

2.2.4 Transformação

O processo de transformação é o que antecede o processo de mineração, e exatamente por isso seu objetivo visa diretamente a otimização dos dados para um melhor resultado da mineração. A transformação dos dados pode ocorrer por meio das atividades de suavização dos dados, adição, agregação, normalização e discretização de atributos.

A suavização é utilizada para remover ruído dos dados, se configurando como uma atividade próxima a de limpeza dos dados, se diferenciando neste caso somente pelo momento em que é aplicado, com sua aplicação podendo ser aplicada por meio de algoritmos de *binning*, regressão e agrupamento.

Já os processos de adição e agregação de atributos de relacionam com o processo de redução dos dados, pois tanto a criação de um atributo através do conteúdo dos outros atributos e a agregação de vários atributos em um são estratégias para reduzir a dimensionalidade.

Por sua vez a normalização de atributos passa pela conversão de unidades de medidas para um mesmo padrão de mensuração, além da conversão dos limites dos atributos para intervalos menores como [-1,1] e [0,1], visando otimizações para a mineração.

Na sequência, a discretização está ligada também ao conceito de redução dos dados, pois visa uma simplificação dos dados originais e tem como resultado uma mineração mais eficiente. A discretização é categorizada de acordo com a sua aplicação, como se é utilizado a informação da classe ou se o funcionamento é ascendente ou descendente (HAN et al., 2012).

2.2.5 Mineração de Dados

A mineração de dados tem sua origem traçada da estatística, pois esse meio já estudava formas de se obter informação dos dados desde antes da existência de computadores, por meio de vários métodos que estão contidos na análise exploratória de dados, que faz parte da mineração de dados (RATNER, 2017).

Esse processo de descoberta de conhecimento é algo que tem obtido grande atenção nos últimos anos devido a crescente quantidade de dados acumulados de diversas fontes, sendo essas informações obtidas definidas como algo que não trivial, que não era conhecido anteriormente e que tem potencial em ser útil (BRAMER, 2016; AGRAWAL et al., 1993).

Por sua vez quanto aos objetivos da mineração, Han et al. (2012) definem que eles podem ser divididos em duas categorias de interesse, interpretação e predição, e apesar de possuírem semelhanças, seus usos são distintos. Entretanto em alguns casos as análises podem unir características de ambos os grupos (AALST, 2016).

Na interpretação a busca é por encontrar padrões nos dados para encontrar regras que possam ser entendidas pelo usuário da aplicação, sendo regras que devem ser originais e acrescer informações que não seriam óbvias, identificando as características do grupo desejado (VERCELLIS, 2009; RATNER, 2017; CHEN et al., 1996).

Isso difere na predição, onde o objetivo é antecipar o valor que as variáveis podem assumir no futuro através dos dados atuais, de acordo com características já vivenciadas em eventos prévios e presentes nos dados utilizados para realizar as previsões, atuando especificamente para cada indivíduo do grupo (VERCELLIS, 2009; RATNER, 2017; KUHN; JOHNSON, 2013).

Leskovec et al. (2014) caracterizam o funcionamento da mineração em duas abordagens, estatística e aprendizado de máquina. Com a primeira abordagem sendo muito usada para casos onde os objetivos são mais diretos, enquanto o aprendizado de máquina é mais utilizado quando o objetivo não é tão claro. Esses campos ainda podem ser utilizados em uma intersecção, o aprendizado estatístico, que é utilizado em diversos problemas como predição de valor ações ou estimativa de glicose para um diabético (HASTIE ROBERT TIBSHIRANI, 2009).

Han et al. (2012) listam diversas formas de se minerar dados, nas análises que buscam interpretar os dados existentes, a forma mais intuitiva para isso é o reconhecimento de padrões frequentes. Um exemplo de sua aplicação é na análise de associação, onde o objetivo é encontrar ligações entre conjuntos, sendo representada com a implicação $X \implies Y$, informando que a

existência do primeiro conjunto implica que provavelmente existirá o segundo, essa implicação possui uma porcentagem de acerto, e é ele que mede a associação (BRAMER, 2016).

Outra forma de análise é por agrupamento, que consiste em agrupar uma série de itens, para que os elementos em um grupo sejam os mais semelhantes possível e os mais diferentes dos elementos dos outros grupos(RATNER, 2017).

Já em análises preditivas destacam-se a classificação e a regressão. A primeira delas é utilizada para atribuir uma classe ao elemento de acordo com as suas características, utilizando para isso regras de classificação, árvores de decisão, fórmulas matemáticas ou redes neurais (HAN et al., 2012). A regressão por outro lado trabalha com estimativa e predição de uma forma numérica, utilizando os dados existentes para criar um modelo matemático que seja capaz de entender os dados atuais e prever os dados futuros (LAROSE; LAROSE, 2014).

Na próxima seção, as técnicas de mineração de dados denominadas agrupamento, associação, classificação e regressão serão apresentadas de forma mais detalhada.

2.2.6 Métodos de Mineração de Dados

O desenvolvimento dos métodos utilizados para mineração de dados foi iniciado pelos matemáticos, em um período prévio a existência de computadores, a partir do desenvolvimento de técnicas da estatística e estando muito conectado com a parte de testes de hipóteses, que é muito utilizada para verificar resultados de uma análise (WITTEN et al., 2016).

Outras formas da estatística utilizadas são os modelos estatísticos, que são importantes pois fornecem um meio matemático para a compreensão e predição de variáveis, que permitem entender um comportamento em um contexto, além do uso da estatística para apresentar os resultados, os tornando mais compreensíveis (HAN et al., 2012).

Já com um maior desenvolvimento das técnicas computacionais tem-se a elaboração de métodos que fazem o uso do aprendizado de máquina, o qual é um campo derivado da inteligência artificial atuando na pesquisa métodos para realizar predições a partir de dados, e se subdivide em dois campos de estudo, aprendizado supervisionado e não supervisionado (WEITZEL et al., 2016).

No aprendizado supervisionado itens rotulados são utilizados para treinar o indutor, já no aprendizado não supervisionado os itens utilizados não devem possuir nenhum rótulo, ainda podendo existir um misto de ambos, que é nomeado aprendizado semi-supervisionado, o qual é muito útil por auxiliar em casos onde a quantidade de dados é pequena (THEODORIDIS;

KOUTROUMBAS, 2009).

2.3 AGRUPAMENTO E SUA APLICAÇÃO EM DADOS GEORREFERENCIADOS

Definido por Leskovec et al. (2014) como o processo de analisar um conjunto de itens e de alguma forma os agruparem, de forma que cada grupo só tenha elementos com características similares entre si e bem distintos dos elementos de outro grupo, como no exemplo da Figura 6, onde as raças de cachorro *beagle*, *daschund* e *chihuahua* são diferenciadas utilizando os atributos de altura e peso, formando grupos que identificam cada uma das raças.

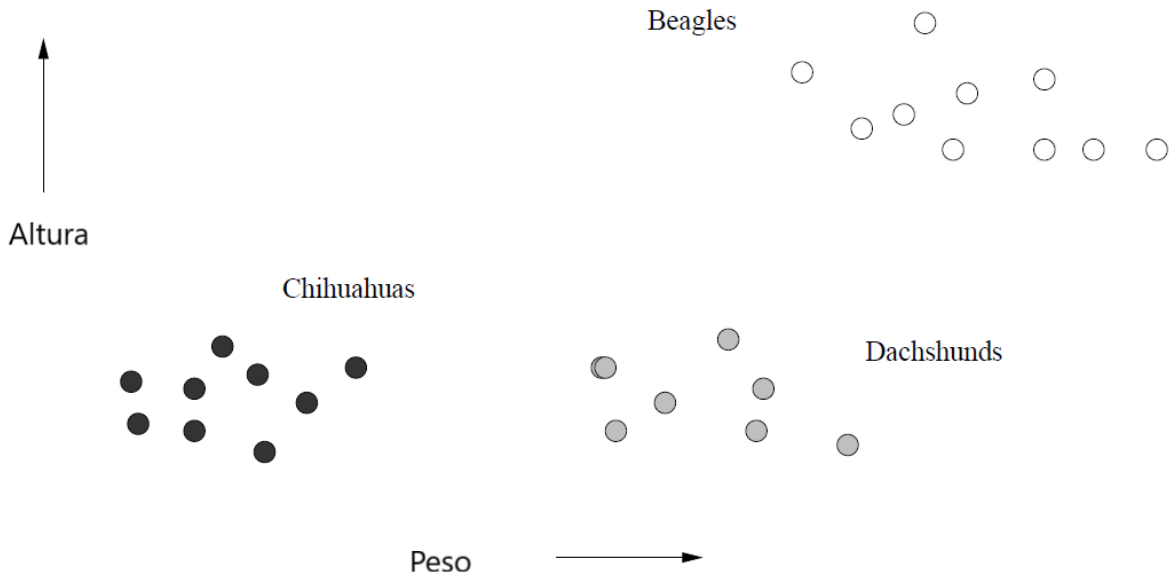

Figura 6 – Exemplo de agrupamento.
Fonte: Adaptado de Leskovec et al. (2014)

Dentre os diversos usos para esse método Han et al. (2012) citam *Business Intelligence*, visando entender perfis dentro de um grande número de clientes; reconhecimento de padrões em imagens, para o identificar tipos de caligrafia para realizar o reconhecimento de um texto escrito a mão; buscas na Internet, para apresentar os resultados de uma forma mais concisa e compreensível; além de usos na biologia e segurança.

Outra aplicação muito interessante para o agrupamento é de forma conjunta com o uso de SIG. Essa abordagem vem ganhando destaque devido a grande variedade, profundidade e quantidade dos dados, enquanto o fato de a estrutura dos dados ser mais complexa vem

motivando a busca por novos métodos (LI et al., 2015).

Dentro das características dos métodos de agrupamento, existem as que diferem no funcionamento. O mais simples deles é quando o espaço multidimensional é dividido para mostrar cada grupo, já outros algoritmos utilizam uma abordagem que aceita que uma instância faça parte de mais de um grupo, outros realizam o agrupamento através de cálculos probabilísticos, e ainda existem algoritmos que trabalham de forma hierárquica, onde grandes grupos possuem subgrupos internamente, todas essas características são apresentadas melhor na Figura 7, com o item A apresentando um exemplo de saída de uma demonstração de como as instâncias foram atribuídas aos grupos; o item B apresentando um exemplo de algoritmo que aceita que instâncias pertençam a mais de um grupo; item C um exemplo de algoritmo que apresenta as probabilidades de um item pertencer a cada grupo; e o item D que apresenta um exemplo de agrupamento hierárquico (WITTEN et al., 2016).

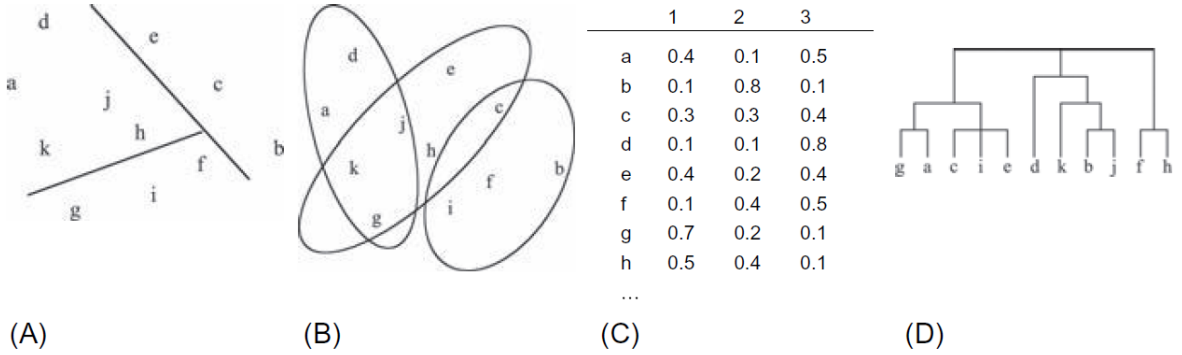

Figura 7 – Exemplo de tipos de agrupamento.

Fonte: (WITTEN et al., 2016)

Entretanto existem alguns detalhes que devem ser considerados na utilização desse algoritmo, esses são a flexibilidade, pois muitas implementações de algoritmos só são capazes de mensurar dados numéricos; a robustez, já que é importante que o método lide bem com ruídos nos dados; e a eficiência, pois a competência em criar grupos com grande quantidade de dados, algo que pode afetar a robustez (VERCELLIS, 2009). Os métodos de agrupamento estão inclusos na já mencionada divisão entre a abordagem estatística ou de aprendizado de máquina. A abordagem estatística já foi muito explorada, dando origem aos algoritmos *k-means*, *k-medoids*, que surgiram com outro objetivo de aplicação, e adquirindo aspectos que o relacionam mais ao aprendizado de máquina quando posteriormente adaptado para a aplicação para mineração de dados. Já a abordagem por aprendizado de máquina é mais recente, e tem como foco o agrupamento com grande quantidade de dados, e como abordagem de aprendizado a utilizada é o não supervisionado.

Segundo Larose e Larose (2014) os algoritmos de agrupamento podem ser divididos em duas classificações, hierárquicos e não-hierárquicos:

2.3.1 Hierárquico

Os métodos hierárquicos funcionam com base na decomposição do conjunto de dados, podendo ainda ser trabalhado de forma aglomerativa ou divisiva, enquanto a primeira tem a característica *bottom-up*, ou seja, constrói o agrupamento dos menores grupos para os maiores, o segundo possuí as características *top-down*, iniciando com os grupos mais abrangentes. Para realizar o agrupamento pode ser utilizado abordagens com base em distância, densidade e continuidade (HAN et al., 2012).

Na Figura 8 é possível visualizar o resultado de um algoritmo hierárquico construído de forma aglomerativa, nela fica claro como os grupos podem ser constituídos a partir de elementos isolados ou de outros grupos.

2.3.2 Não-Hierárquicos

Já os algoritmos não-hierárquicos possuem diferentes características, mas a maior parte utiliza a abordagem interativa que funciona a partir da especificação da quantidade de grupos desejados, realizando assim a divisão dos grupos por particionamento, podendo ser esse também um dos termos para se referir a essa categoria de técnicas.

Assim a partir da definição de k, são definidos itens chave, que são os itens que serão a base do agrupamento, essa definição pode ocorrer utilizando diversas técnicas que vão afetar o resultado final. Os outros itens têm seu grupo definido de acordo com a sua distância para os itens chave. O principal exemplo dessa abordagem é o algoritmo *k-means* (LI et al., 2015).

Na Figura 9 é demonstrado como um item é selecionado para um grupo de acordo com

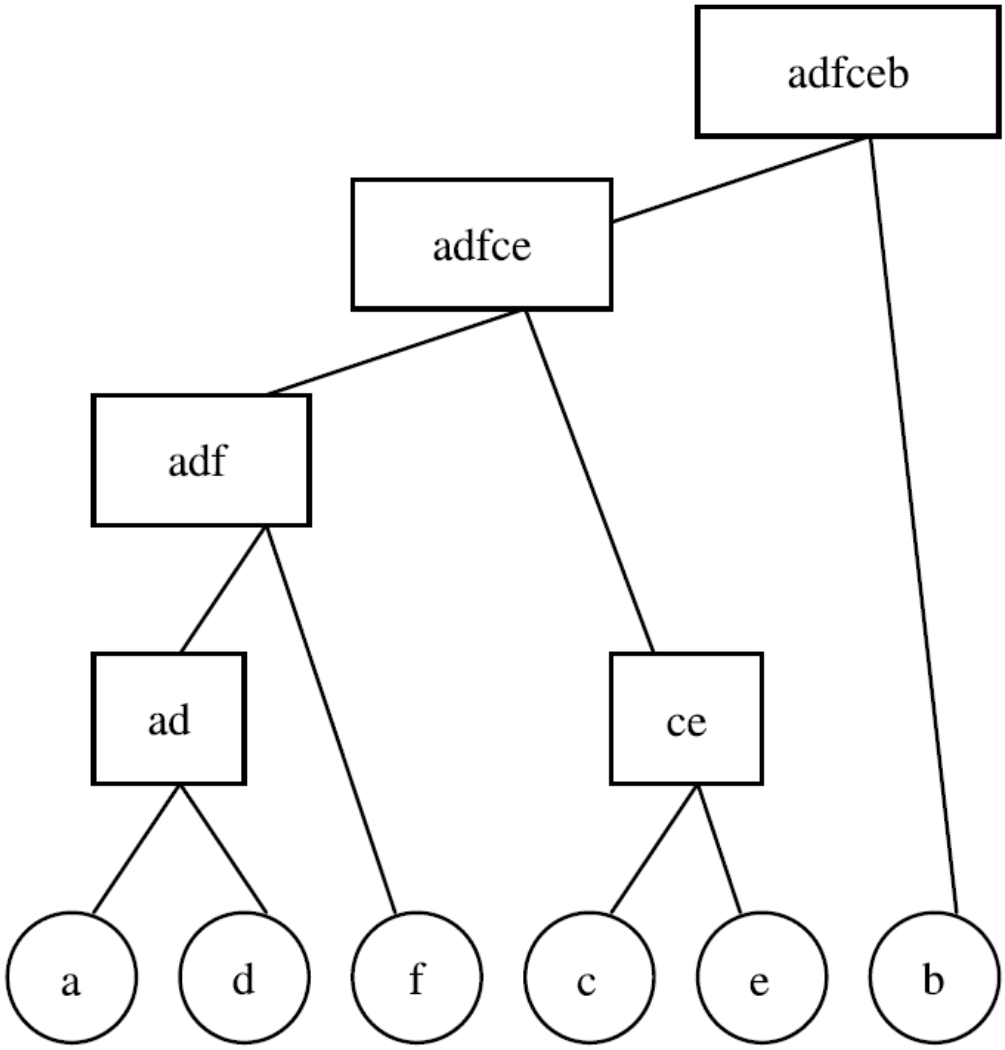

Figura 8 – Exemplo de agrupamento hierárquico.

Fonte: (BRAMER, 2016)

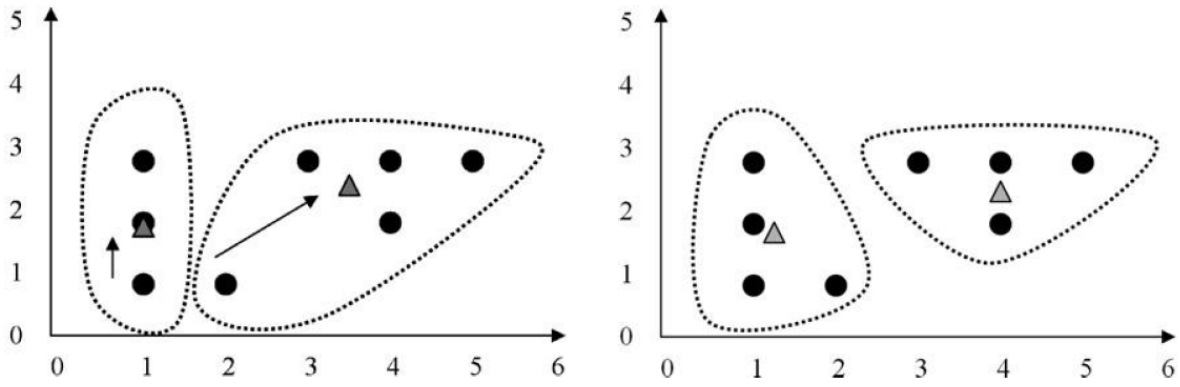

Figura 9 – Exemplo de agrupamento por particionamento.

Fonte: (LAROSE; LAROSE, 2014)

a proximidade com o item central.

2.3.3 Distâncias

Os algoritmos de agrupamento utilizam meios de mensuração, o mais básico deles é a distância. Existem diversas distâncias que podem ser utilizadas, a mais popular delas é a distância Euclidiana, mas ainda é possível citar a Chebyshev, Manhattan e Minkowski (que é uma generalização de todas as anteriores) (LI et al., 2015).

A distância Euclidiana é dada pela seguinte equação:

$$de = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 \cdots (x_n - y_n)^2} \tag{1}$$

Onde cada um dos elementos x_1 até x_n representa um item de um vetor X, da mesma forma que com os elementos de y_1 até y_n são parte de um vetor Y. Ambos os vetores X e Y, representam as instâncias entre as quais é feito o cálculo da distância, com o intervalo entre 1 e n representando o número de dimensões em que essas instâncias são representadas e que serão consideradas para o cálculo da distância.

Por sua vez a distância Manhattan é dada pela seguinte equação:

$$dma = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

Os elementos x e y utilizados nesse cálculo são os mesmos utilizados anteriormente.

Já a distância Minkowski é dada pela seguinte equação:

$$dmi = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \tag{3}$$

Os elementos x e y são da mesma forma que na distância euclidiana, representações de itens dos vetores X e Y respectivamente, enquanto p é um valor definido pelo usuário, o que pode ocorrer através da inequação de Minkowski, e deixa evidente como a distância Euclidiana é um exemplo com p adquirindo o valor 2 e a Manhattan quando adquire o valor 1.

Por fim a distância Chebyshev é dada pela seguinte equação:

$$dc = \max_i |x_i - y_i| \tag{4}$$

Os elementos utilizados são novamente os já citados nas distâncias anteriores, com a variável i assumindo todos os valores possíveis entre as dimensões dos vetores X e Y.

Grande parte das implementações dos algoritmos de agrupamento lida apenas com um número limitado de distâncias, sendo que a maioria deles não é capaz de lidar com distâncias mais complexas como a que envolve a curvatura terrestre.

2.3.4 Densidade

Já na mensuração através da densidade pode ser aplicada da mesma forma tanto a abordagem hierárquica quanto com a não-hierárquica. Elas funcionam utilizando a metodologia de realizar a seleção dos itens para o agrupamento de acordo com o valor da densidade mínima desejada.

Essa densidade é calculada a cada item verificado, sempre analisando se em caso da adição do item ao grupo a densidade se mantêm como desejado, essa densidade muitas vezes é calculada a partir de uma métrica de distância, podendo essa ser uma das apresentadas aqui ou outra qualquer. Dentre as coisas que essa abordagem permite é a formação de grupos sem uma forma específica, pois a formação dos grupos ocorre de forma arbitrária (HAN et al., 2012).

Na Figura 10 é mostrado por meio da execução do algoritmo DBSCAN como o agrupamento por densidade seleciona os itens de acordo com a proximidade dos vizinhos, utilizando a densidade desejada por meio do valor definido de épsilon como parâmetro para definir se um item vai ser incluso no grupo. Na execução do algoritmo os pontos o e p foram definidos como os pontos iniciais, e por meio da verificação de densidade mínima os pontos m e q foram selecionados para o grupo de p, enquanto os pontos s e r foram selecionados para o grupo de o, sendo que para a seleção de r foi necessário que um caminho de pontos fosse selecionado.

2.3.5 Aplicação em Dados Georreferenciados

Com o desenvolvimento dos itens apresentados nesta seção e na 2.1, passa a ser possível o estudo da área de intersecção entre esses conhecimentos. Além disso a grande quantidade de dados disponíveis permitiu que em 1995, Koperski e Han (1995) abordassem essa possibilidade como algo viável.

Segundo Laurini (2017) contexto de dados com informações geográficas vem se

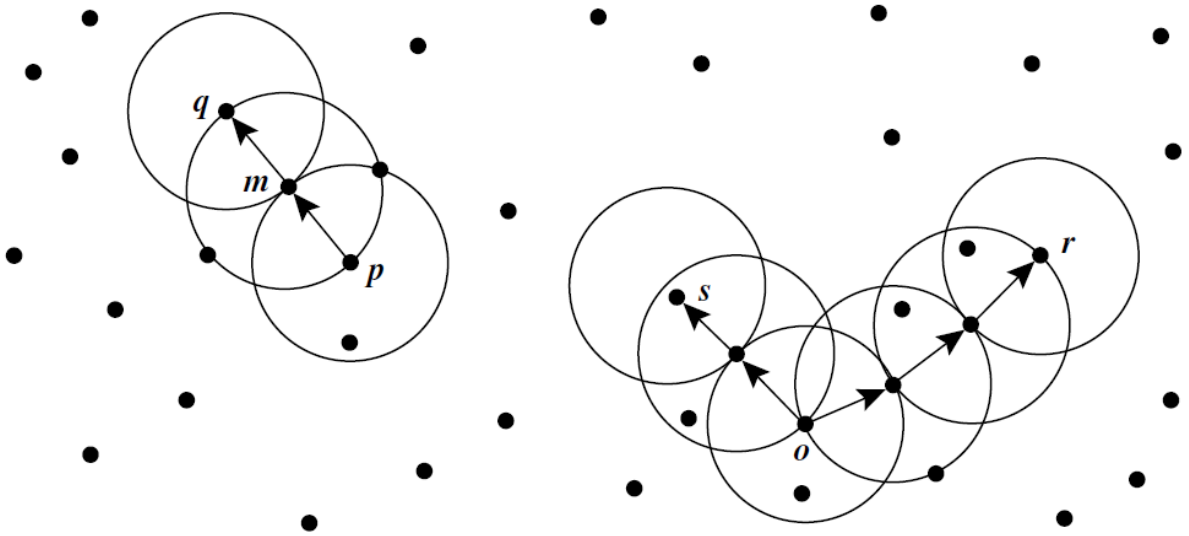

Figura 10 – Exemplo de agrupamento por densidade.
Fonte: (HAN et al., 2012)

mostrando interessante para a exploração com uso da mineração de dados, pois o conteúdo é mais rico em profundidade e largura que dados tradicionais, com tipos variados de informação, que ainda tem crescido muito em volume (LI et al., 2015). Como alguns exemplos práticos dessa mineração Zheng (2015) cita as várias possibilidades de descoberta sobre as trajetórias e Huang et al. (2004) citam a descoberta de padrões de hospedagem de servidores por meio de dados espaciais, ambos fazendo uso dessa intersecção de conteúdos.

Segundo Han et al. (2012) para a realização da mineração de dados podem ser utilizados como fontes tanto os dados vetoriais quanto os *raster*, com a possibilidade da realização de diversos tipos de análise, como agrupamento, associação e classificação, que já foram apresentados na Seção 2.2.

Entretanto como já dito na mesma seção, as técnicas de agrupamento se adaptam muito bem ao uso de dados espaciais, pois são capazes de fornecer informações sobre a concentração de itens, e por isso esse será aprofundado, juntamente como exemplos de sua aplicação em dados espaciais (LI et al., 2015).

2.4 CLASSIFICAÇÃO E SUA APLICAÇÃO EM TEXTO

Segundo Witten et al. (2016) a classificação é o processo de a partir de classes pré-definidas rotular itens sem uma classe atribuída, por meio da criação de uma regra para

classificar, sendo assim um algoritmo preditivo, e que, portanto, apresenta erro. Para realizar essa classificação o algoritmo é dividido em duas etapas, a primeira de aprendizado, onde a função de classificação é criada, e uma segunda etapa onde ocorre a classificação dos itens com classes desconhecidas, aplicando a função anteriormente criada. Esses processos podem ser visualizados respectivamente na Figura 11 (HAN et al., 2012).

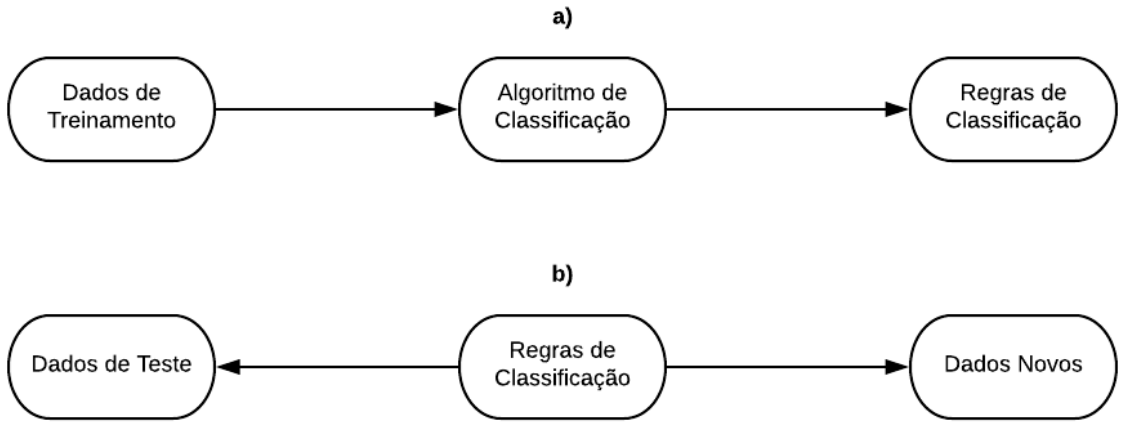

Figura 11 – Exemplo de das duas etapas do processo de classificação.
Fonte: Adaptado de Han et al. (2012)

Para se criar a função de classificação Bramer (2016) cita os métodos Nearest Neighbour Matching, que tenta realizar a classificação identificando a proximidade da instância não classificada com as já rotuladas; Regras de Classificação, que basicamente criam condições para realizar a classificação e Árvores de Decisão, que criam uma estrutura de classificação seguindo condições.

Han et al. (2012) ainda citam a possibilidade da aplicação de Redes Neurais Artificiais na classificação, que utilizam uma coleção de neurônios artificias conectados para atribuir as classes, outras possibilidades para a classificação são o uso de classificação bayesiana e máquinas de suporte vetorial. Na Figura 12 pode-se visualizar exemplos de classificação com Regras de Classificação, Árvores de Decisão e Redes Neurais respectivamente.

A classificação pode ser utilizada em muitas aplicações, como verificação de fraudes com cartões de crédito; posicionamento de um estudante com necessidades especiais em um serviço de transporte; indicar riscos de aplicações financeiras; diagnósticos médicos; verificação de fraude em assinaturas e verificação de compras fora do padrão do cliente (LAROSE; LAROSE, 2014).

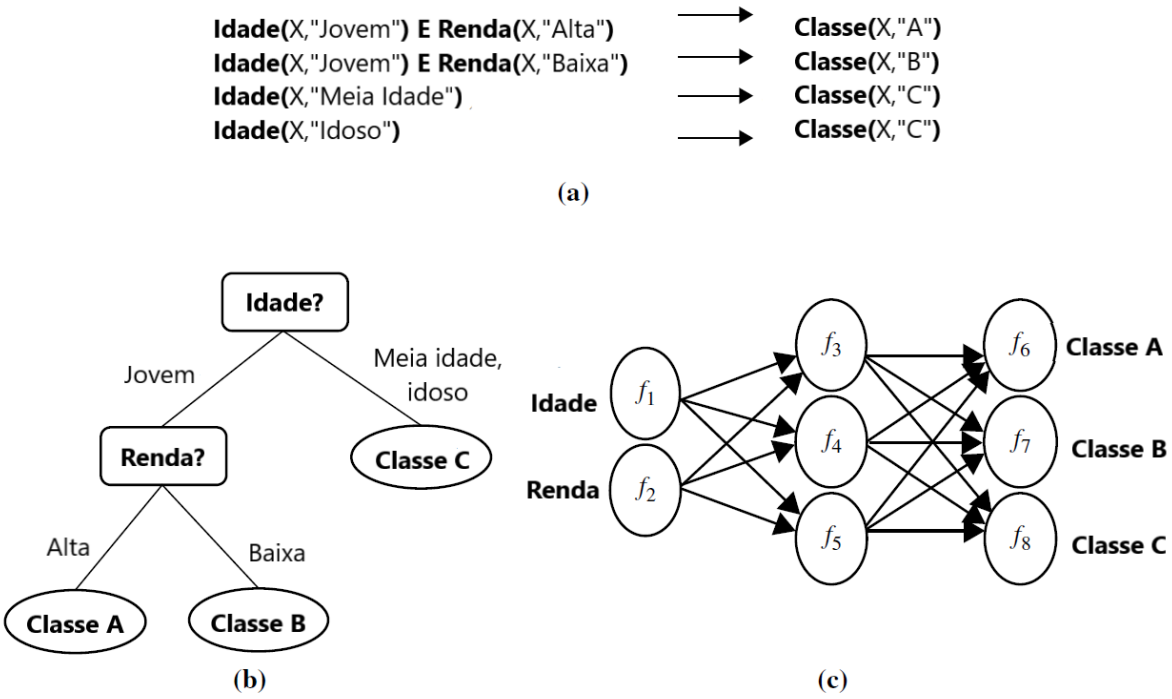

Figura 12 – Exemplo de tipos de métodos de classificação.

Fonte: Adaptado de Han et al. (2012)

2.4.1 Determinação da classe

O funcionamento de um classificador via de regra utiliza o cálculo da probabilidade de um item ser de determinada classe, com geralmente essa saída sendo entre o intervalo de 0 e 1, e enquanto alguns classificadores tem sua saída em uma definição mais estatística onde a soma de todos os itens é igual a 1, em alguns classificadores como rede neural a saída não necessariamente terá a some de todos os itens como 1, sendo nesses casos aplicada uma função de transformação para que os valores sejam adequados a uma saída propriamente probabilística (KUHN; JOHNSON, 2013).

Com os resultados tendo seus valores entre 0 e 1, se faz necessário o uso de uma determinada métrica para a atribuição das classes de acordo com a probabilidade. O pensamento mais óbvio inicialmente é de utilizar a maior das probabilidades como a resposta para qual classe será atribuída, entretanto nem sempre esse é o procedimento mais interessante, e por isso pode-se fazer o uso de uma função de custo para a determinação, existindo diversas métricas de função que podem ser escolhidas e ainda diferentes técnicas para se obter os melhores parâmetros para a função (PACHECO et al., 2016).

2.4.2 Aplicação em Texto

Uma das derivações da mineração de dados, já apresentada na Seção 2.2, é a mineração de texto, que busca obter informações desse tipo de dado. Entretanto a análise de textos tem muitas diferenças em relação a abordagem com dados numéricos, pois os textos não são organizados de forma tão estruturada (WEISS et al., 2005).

Miner (2012) data o início do processo de extrair informações de texto em 1987, por meio das Conferências de Compreensão de Mensagens, que analisaram mensagens militares e mensagens relacionadas a terrorismo na América Latina.

Para lidar com a ausência de uma estrutura definida, a mineração de texto utiliza conhecimentos interdisciplinares como o aprendizado de máquina, a estatística e o Processamento de Língua Natural (NLP). O procedimento de mineração de textos normalmente pode incluir classificação de texto, agrupamento de texto, análise de sentimento e várias outras atividades, normalmente executando somente uma delas (HAN et al., 2012).

Se tratando especificamente da análise de sentimentos, ela é definida como o processo de identificar o pensamento do autor por meio do texto, tendo grande aplicabilidade em dados disponíveis na Internet, principalmente *blogs* e redes sociais, pois os mesmos possuem grande quantidade de opiniões (MINER, 2012).

Esse tipo de análise enfrenta vários desafios na sua execução, muitos deles oriundos do NLP e da comunicação, além das questões relacionadas a profundidade, que passa pela classificação de um documento como um todo, de cada frase ou ainda de cada aspecto (LIU, 2015).

Para a identificação de sentimentos existem abordagens que partem da parte de análise léxica utilizando para isso características estatísticas, ou por aprendizado de máquina utilizando características dos classificadores dessa abordagem (POZZI et al., 2016).

2.4.3 Abordagem Léxica

A abordagem léxica é definida como a identificação da orientação semântica do texto de acordo com o uso de um conjunto específico do vocabulário, sendo uma abordagem muito utilizada em pesquisas acadêmicas (YANG; MO, 2016). Dentro dessa abordagem, Weitzel et al. (2016) definem duas subdivisões para abordar o problema, baseado em um dicionário e em um

Corpus com análise sintática e semântica. Para a primeira abordagem, é preciso que se tenha os dicionários, e para isso existem dois meios de obtenção, a criação manual ou a construção automática (TABOADA et al., 2011).

Como método para aplicação de análise de sentimentos léxica com base em dicionário, Farhadloo e Rolland (2016) cita o método *Stemming*, que busca analisar palavras de acordo com o seu radical, o vocabulário controlado, que busca fazer a análise a partir de um dicionário restrito a palavras pré-determinadas, e o Análise Semântica Latente (LSA), que realiza a análise buscando identificar a relação entre as palavras e a frequência que as mesmas têm no texto.

Já para a análise baseada em um Corpus, a abordagem é que o classificador deve aprender com o Corpus, para então saber analisar os textos, sendo uma abordagem de aprendizado semi-supervisionado, enquanto a abordagem por dicionário é não supervisionado (BENEDETTO; TEDESCHI, 2016).

Entretanto a abordagem léxica possui alguns problemas, como a dificuldade em obter informação de uma frase quando o atributo está de forma oblíqua na frase, algo comum na construção de sentenças (HU; LIU, 2004). Existem diferentes técnicas para abordar esse problema, sendo a grande parte delas baseada em um Corpus para realizar a identificação do atributo, com Fei et al. (2012) propondo uma abordagem de dicionário para maximizar a quantidade de atributos encontrados.

2.4.4 Aprendizado de Máquina

Já para a classificação com o aprendizado de máquina, os algoritmos utilizam as possibilidades já apresentadas como árvores de decisão e redes neurais artificiais.

Para realizar a classificação os algoritmos tendem a utilizar características chave para o texto, tendo abordagens diferentes para isso, Liu (2015) cita a análise de termos que aparecem e com que frequência isso ocorre; análise de partes do discurso e a relação de palavras próximas; análise dos sentimentos das palavras ou frases; uso de regras de opinião; análise de palavras que alteram o sentimento da frase e análise de dependência sintática.

Já Pozzi et al. (2016) cita que a análise também pode ser feita por meio de conteúdo paralinguístico, já que em alguns casos como redes sociais o uso de *emoticons*, a ênfase inicial, onomatopeias, alongamento de palavras, letras maiúsculas e *hashtags*.

Os métodos de aprendizado de máquina mais utilizados para a classificação de texto são os baseados em regras de classificação, bayesianos e Máquinas de Vetores de Suporte

(SVM), sendo que cada um deles realiza uma abordagem diferente do problema (AGGARWAL, 2014).

Enquanto a abordagem bayesiana utiliza uma forma de aprendizado probabilístico, as regras de classificação utilizam uma abordagem de se-então, muitas vezes utilizando árvores de decisão para representar essas condições (WITTEN et al., 2016). Já a SVM é uma técnica de reconhecimento de padrões que deriva do conceito de classificadores de margem (AGGARWAL, 2014).

2.4.5 Análise de Sentimentos em Redes Sociais

As redes sociais se popularizaram muito com os avanços da Internet, essa popularização tem como consequência a grande presença de dados sobre muitas pessoas, o que permite a obtenção de diversas informações sobre as pessoas. Entretanto apesar de sociólogos estarem estudando as conexões das pessoas a mais de um século, somente com o advento das redes sociais que se passou a contar com uma grande quantidade de dados sobre as relações sociais, implicando na necessidade de novas formas de se analisar isso (AGGARWAL; ZHAI, 2012).

Um dos meios de realizar essa análise é pela análise de sentimentos, nela após a obtenção dos dados de redes sociais, são aplicados os métodos para tentar extrair informações de sentimentos presentes na rede social, transformando o conteúdo em informação, que pode ser muito útil em diversos casos de tomada de decisão (WEITZEL et al., 2016).

Para a análise de sentimentos em redes sociais é preciso levar em consideração alguns fatores sobre as características das mesmas, como o fato de as mensagens serem mais curtas, apresentarem muito ruído, possuírem um grande dinamismo temporal, a existência de informação implícita, o uso de várias linguagens e as relações presentes no texto (POZZI et al., 2016).

Considerando esses fatores existem duas abordagens de informações que podem ser obtidas, enquanto uma delas foca na análise da linguagem utilizada, obtendo informações das características textuais, a outra analisa principalmente as relações expostas no texto, buscando entender seu status e valor, como representado na Figura 13, com a análise sobre a linguagem natural se dividindo para análises sobre o texto, parte do discurso e o conteúdo paralinguístico, se relacionando com o conteúdo em si; advérbios e adjetivos; e *emojis* e *hashtags* respectivamente. Enquanto a análise sobre as relações se divide em *status* homofílico

e valor homofílico, sendo que o primeiro é baseado em informal, formal ou categoria atribuída, enquanto o segundo se baseia em valores atitudes e crenças.

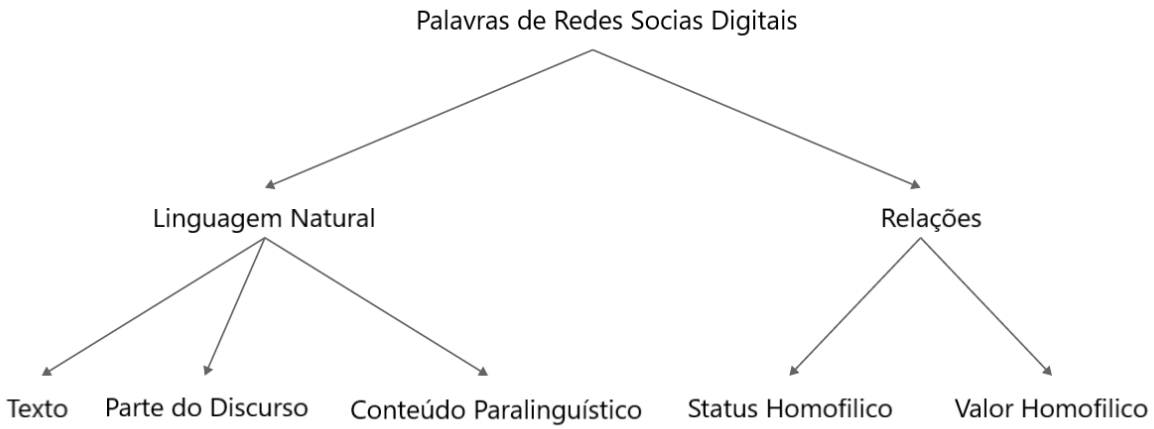

Figura 13 – Subdivisões da Análise de Sentimentos em Redes Sociais.
Fonte: Adaptado de Pozzi et al. (2016)

2.4.6 Identificação de Relação com a Frase

A análise de sentimentos pode ser utilizada para identificar o tipo de texto, como se o texto é uma opinião do autor ou uma notícia, ou mesmo na tentativa de identificação de piadas ou ironias.

Para isso é possível utilizar tanto a abordagem léxica como feito no trabalho de (GOMIDE et al., 2011) ou de aprendizado de máquina como no trabalho de (JI et al., 2015), em ambos os casos análise foi feita com dados de redes sociais, que se mostram um ambiente propício para isso.

2.5 TRABALHOS CORRELATOS

2.5.1 Trabalhos com mineração de dados georreferenciados aplicada em rede social

Em relação aos trabalhos que buscam realizar a mineração de dados georreferenciados em redes sociais, Stock (2018) selecionou 690 trabalhos relevantes de diversas fontes, identificando que desses trabalhos 54,2% utilizavam como fonte dos dados o Twitter, sendo que 8,1% dos trabalhos aplicados a identificação de problemas relacionados a saúde, com 0,5% aplicados a dengue.

Alguns dos trabalhos de maior destaque que fazer uso da rede social Twitter para realizar a mineração de dados georreferenciados são os de Longueville et al. (2009), Boulos et al. (2011), Albuquerque et al. (2015), Shelton et al. (2015), Steiger et al. (2015), Longley et al. (2015) os quais fazem desde uma análise demográfica, passando por uma análise do planejamento urbano, até análises para identificação de catástrofes.

2.5.2 Trabalhos com mineração de dados georreferenciados aplicada em rede social aplicados a saúde

Já no que tange a aplicação em casos de saúde, se tem os já citados trabalhos de Ghosh e Guha (2013), Allen et al. (2016), mas ainda se pode destacar os trabalhos de Dredze et al. (2013), Yang e Mu (2015) que respectivamente abordam a identificação de casos de gripe e a identificação de casos de depressão entre os usuários do Twitter.

Como trabalhos aplicados a dengue destacam-se Gomide et al. (2011) e Sousa et al. (2018), ambos realizaram a análise por meio do Twitter, realizando também a classificação dos *tweets* para identificar os casos de dengue, mas diferindo na quantidade de classes e na metodologia para isso. Para a identificação geográfica ambos usaram os *tweets* georreferenciados, sendo que Gomide et al. (2011) realizou a aplicação do agrupamento por densidade para identificar as regiões e Sousa et al. (2018) não, e enquanto o primeiro realizou

uma comparação de dados oficiais de algumas cidades brasileiras com os dados geográficos, o segundo realizou uma apresentação visual da concentração.

3 MATERIAIS E MÉTODOS

Nesse capítulo serão apresentados todos os materiais utilizados nesse trabalho e todos os métodos que serão aplicados no mesmo.

3.1 MATERIAIS

3.1.1 Dados do Twitter

O Twitter foi definido como a rede social que foi utilizada nesse trabalho por uma série de motivos como: a já citada grande utilização entre os trabalhos recentes; a comunicação majoritariamente através de texto; a facilidade de obtenção dos dados, já que o Twitter disponibiliza uma API para isso.

A API disponibilizada pelo Twitter requer a submissão de um projeto, que é analisado se está de acordo com as políticas da empresa, em caso de o projeto ser aprovado o usuário recebe autorização para fazer as requisições pedidas.

Cada modelo de requisição possui determinadas características, podendo ser uma captura de *tweets* em tempo real ou buscas por conteúdo, que podem atingir *tweets* nos últimos 7 ou 30 dias, ou ainda em todo o período de atividade do Twitter. Além das datas essa busca permite definir como parâmetros palavras chave que sejam interessantes e limites de localização por meio de coordenadas geográficas. Assim é possível definir buscas por dengue e que estejam no estado do Paraná.

No caso da aplicação desse trabalho a abordagem escolhida para a obtenção dos dados para os testes e a elaboração do classificador foi a busca mista, fazendo uso tanto da busca em

todo o histórico quanto a busca limitada por 30 dias, de forma a tentar maximizar a quantidade de *tweets* obtidos para a elaboração do classificador.

Já a obtenção dos dados para a realização da análise comparativa, foi definido o uso da busca em todo o histórico, delimitando-a a um período de 3 anos, pois esse foi o limite considerado mais interessante, devido ao quantidade máxima de requisições que considerassem a questão geográfica que puderam ser compradas.

Após a definição do modelo é preciso entender como utilizar os dados fornecidos, a API fornece os dados em um formato de Notação de Objetos *JavaScript* JSON, onde as características do *tweet* são informadas da forma demonstrada no Anexo A.

Em relação aos dados geográficos existem diversas formas de se obter essa informação do tweet: a mais precisa delas é a obtenção por meio da localização exata que o usuário escolhe informar; a segunda é por meio de um lugar informado pelo usuário, quando ele insere em seu tweet que está em determinado lugar; a última e menos precisa é a inferência através do local determinado pelo usuário como seu lugar de residência. A forma como são representadas as informações geográficas dos *tweets* pode ser vistas no Anexo B.

3.1.2 Dados de Casos Reais

Para que os resultados obtidos através da análise pelo Twitter possam ter sua eficácia verificada é preciso possuir os dados sobre os casos reais, com o maior nível de precisão possível.

Para isso foi obtida a base de dados do sistema InfoDengue[1], essa base de dados possui os números de casos de dengue de todas as cidades paranaenses em cada semana, com algumas cidades possuindo dados desde 2008, e para a execução do trabalho foram selecionados os dados no período entre 2016 e 2018, como pode ser visualizado na Figura 14, que representa o

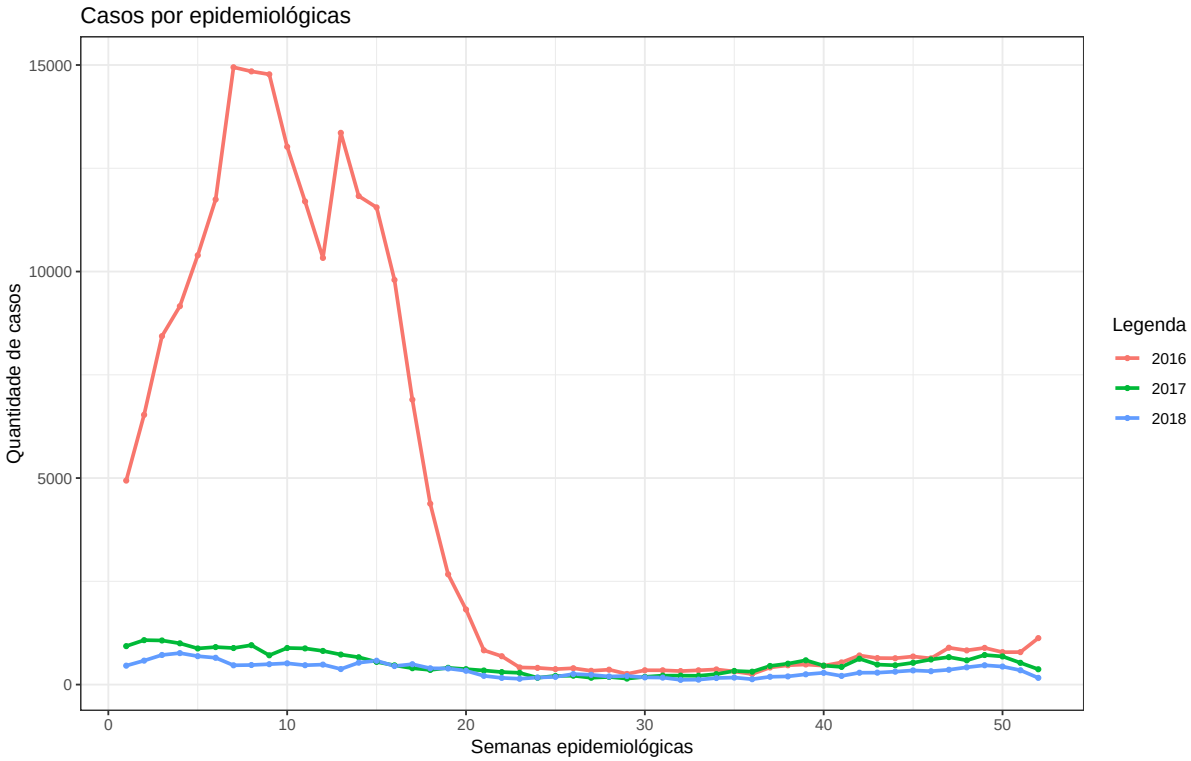

Figura 14 – Dados de dengue no Paraná por cada semana entre 2016 e 2018 segundo o sistema InfoDengue.

Fonte: Autoria Própria

gráfico com os casos por semana em cada um dos anos.

3.1.3 Linguagem de Programação

Para manipular esses dados é necessário o uso de uma linguagem de programação, para isso foi escolhida a linguagem Python[2], pois entre muitas vantagens se destacam a sua simplicidade e a facilidade de manipulação de dados, fatores definidos como fundamentais na execução desse trabalho.

A linguagem além do uso na manipulação dos dados pode ser aplicada na implementação de classificadores e de etapas do pré-processamento, pois apresenta uma grande quantidade de bibliotecas disponibilizadas que facilitam esse processo de implementação. No caso da execução do trabalho foi utilizado para o pré-processamento.

3.1.4 Banco de Dados

Para o armazenamento dos dados é necessário o uso de um Sistema Gerenciador de Banco de Dados (SGBD) que suporte os dados geográficos, nesse ponto o PostgreSQL[3], um SGBD relacional, na versão 11 foi escolhido para o trabalho, pois além de ser um banco de dados gratuito possui a confiabilidade e robustez desejada para o trabalho.

Sua extensão geográfica (PostGIS[4]) em sua versão 2.5, permite que os dados georreferenciados sejam armazenados e manipulados através de cerca de 1500 funções disponibilizadas, que facilitam a execução de diversas tarefas.

[1] info.dengue.mat.br
[2] python.org
[3] postgresql.org
[4] postgis.net

3.1.5 Waikato Environment for Knowledge Analysis

Waikato Environment for Knowledge Analysis[5] (WEKA) é um projeto desenvolvido inicialmente pela universidade de Waikato em 1992 na Nova Zelândia, visando a disponibilização de forma mais acessível de técnicas de aprendizado de máquina através de um software de código aberto.

Nesse trabalho foi utilizado o software em sua versão 3.8, utilizando os métodos classificadores e de pré-processamento que já estão implementados na ferramenta, o que justifica a escolha do uso da mesma, além da sua possibilidade de livre acesso.

3.1.6 NLTK

Como uma outra alternativa para a realização do pré processamento de texto, foi utilizada a ferramenta NLTK[6], a qual segundo os responsáveis pela sua manutenção é uma das plataformas mais importantes construídas em Python para o desenvolvimento de trabalhos que englobem dados de linguagem humana, além de ser distribuída em um projeto de código aberto, o que facilita seu uso no projeto.

A ferramenta possui implementada uma série de algoritmos para a realização do pré-processamento de texto, os quais incluem tolkenização, *stemming*, *tagging*, *parsing* e reconhecimento semântico.

3.1.7 R

R[7] é um ambiente de desenvolvimento e linguagem de programação para métodos estatísticos computacionais e gráficos, estando sobre a licença de um software livre, com a sua versão 3.6 sendo a mais recente no momento.

A escolha da ferramenta ocorreu para a utilização do algoritmo de agrupamento, pois o

[5] cs.waikato.ac.nz/ml/weka/index.html
[6] nltk.org
[7] r-project.org

método selecionado já está implementado na ferramenta, além de características fundamentais como a confiabilidade e robustez do software.

3.2 MÉTODOS

3.2.1 Amostragem

Como métodos de amostragem foram utilizados os métodos Resample e SpreadSubsample, sendo ambos os métodos disponíveis na ferramenta Weka, sendo que em ambas as abordagens é produzida uma amostragem aleatória dos dados, apenas com a diferença de que enquanto a abordagem do método Resample inclui reposição a abordagem do método SpreadSubsample não a considera, tanto os dois métodos, quanto os conceitos de suas amostragens, podem ser vistos de forma mais detalhada no trabalho de Witten et al. (2011).

3.2.2 Pré-processamento de texto

Para a realização do pré-processamento do texto duas abordagens foram utilizadas, uma utilizando os algoritmos implementados pela ferramenta Weka e outra abordagem que considerou o uso da ferramenta NLTK, sendo que na segunda ainda foi feito uma análise sobre a aplicação de condição de frequência mínima de palavras e *stemming*.

Em ambos os casos foi utilizada a abordagem de *bag-of-words* para a transformação do conteúdo presente nos *tweets* para atributos que possam ser compreendidos pelos métodos classificadores, facilitando assim a criação do conjunto de atributos a ser interpretado pelos algoritmos, as características deste método foram aprofundadas no trabalho de Guzella e Caminhas (2009).

3.2.3 Classificação de Texto

Para a realização da classificação de texto foram utilizados os *tweets* obtidos por meio da API, os quais foram rotulados manualmente para que o classificar pudesse ser avaliado adequadamente.

Como métodos para realizar a classificação de texto foram selecionados os utilizados no trabalho Ji et al. (2015), onde é explicada a abordagem do uso de mais de um classificador para identificar caso um *tweet* é uma notícia ou um texto onde o autor tem uma relação direta com o texto. Por isso os classificadores Bayesiano Ingenuo Multinomial e Máxima Entropia foram selecionados para o trabalho.

Já pelo trabalho de Aluísio et al. (2016) os classificadores SVM e J48 foram adicionados ao trabalho, por apresentar um melhor desempenho na classificação de itens na língua portuguesa.

Esses métodos foram avaliados de forma a identificar os grupos em que possuem mais acertos e nos que possuem mais erros, para que possam atuar de modo que se complementem.

3.2.4 Agrupamento de Dados Georreferenciados

Para realizar o agrupamento de dados geográficos a abordagem escolhida foi a por densidade, nessa abordagem o método com maior uso é o DBSCAN. O algoritmo foi publicado na conferencia de KDD em 1996 e vem sendo utilizado além de no trabalho de Gomide et al. (2011) em diversos outros trabalhos como os de Shen et al. (2016), Ma et al. (2013), ele é caracterizado pelo uso da distância euclidiana e pelas diferentes possibilidades de uso já utilizadas em todo o seu período de existência.

3.3 FLUXOGRAMA DE EXECUÇÃO DO PROJETO

Na Figura 15 é apresentado como foi planejada a execução do trabalho, demonstrando onde foram utilizados os materiais e métodos para que o todo pudesse ser concluído.

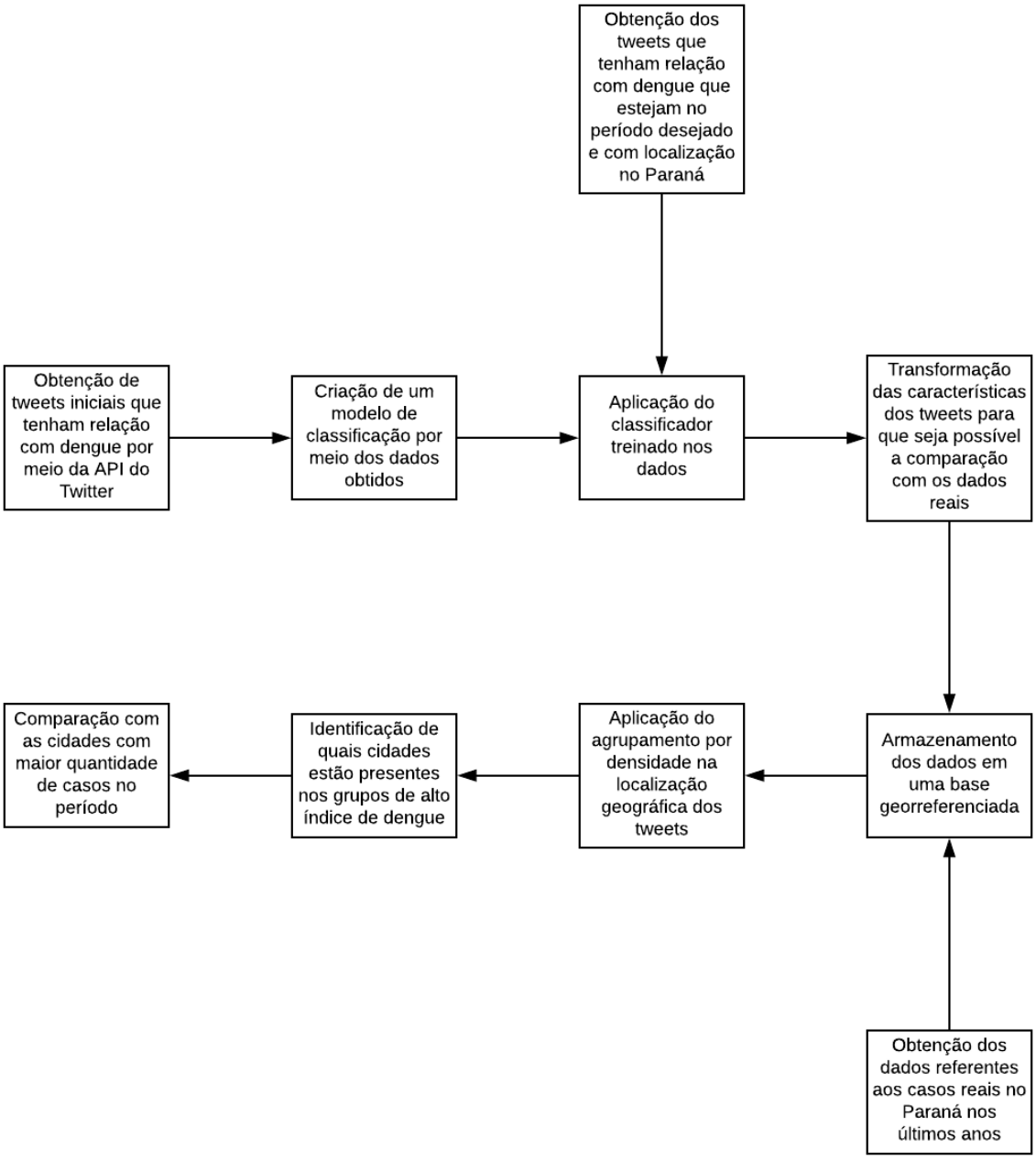

Figura 15 – Fluxograma do Processo de Execução do Projeto.

Fonte: Autoria Própria

Como primeiro passo destacado está a obtenção de *tweets* com a API do Twitter, esses *tweets* foram utilizados para a criação do modelo classificador para identificar quais *tweets* possuem uma relação direta com a dengue, o que está destacado no passo seguinte.

Como próximas etapas estão a obtenção dos *tweets* georreferenciados que estejam na localidade do estado do Paraná e a aplicação do modelo de classificação anteriormente criado, identificando quais *tweets* eram interessantes para a sequência do trabalho.

Após essa identificação algumas características dos *tweets* precisaram ser transformadas, como a data que precisa estar compatível com o intervalo semanal dos dados reais registrados, após isso tanto os *tweets* quanto a base de dados com os registros de casos reais foram armazenados em uma base georreferenciada.

Os dados dos *tweets* foram utilizados como entrada para o método de agrupamento por densidade, que identificou as regiões que pelos *tweets* indicaram maior índice de dengue.

Após isso foi identificado quais são as cidades que estão presentes nessa região e os dados obtidos pelos *tweets* foram comparados com os dados de casos de dengue identificados.

4 RESULTADOS

Neste capítulo serão apresentados os resultados obtidos para a classificação dos *tweets* e para a comparação entre o agrupamento e os dados de casos de dengue reais.

4.1 ESCOLHA DO CLASSIFICADOR

Para realizar a avaliação dos classificadores e desenvolver o modelo que foi utilizado para classificar os *tweets* foi utilizada a API do Twitter para obtenção dos *tweets* que mencionam a palavra dengue. Posteriormente esses *tweets* foram inseridos na base de dados, de forma que o texto dos *tweets* não se repetisse, para evitar a presença de dados repetidos no treinamento do classificador, o que representaria apenas um gasto computacional e não acrescentaria no aprendizado. Esse processo resultou em 5362 *tweets* únicos armazenados na base de dados.

Com os dados podendo ser acessados facilmente através da base de dados, foi iniciado o pré-processamento. Para isso inicialmente foram utilizadas as ferramentas internas do Weka, para realizar as atividades de conversão de tipos, amostragem e transformação do campo de texto em um vetor numérico, o que resultou na presença de 11582 atributos para serem avaliados pelo classificador, para que esses dados pudessem ser utilizados de entrada pelos classificadores escolhidos.

Com o pré-processamento concluído, foi feita a análise dos classificadores utilizando uma divisão do *dataset* de 66% para treino e o restante para teste, utilizando de um formato aleatório para essa divisão, com os testes ocorrendo 10 vezes para cada classificador, para que a consistência pudesse ser aferida. As médias do percentual de acerto de cada um dos classificadores podem ser visualizadas na Tabela 1.

Com o objetivo de otimizar o desempenho dos classificadores, iniciou-se um processo de encontrar possibilidades que permitissem reduzir a dimensionalidade, para tornar o processo computacionalmente menos complexo, sem que houvesse perda de conhecimento. Para isso

Tabela 1 – Percentuais do desempenho dos classificadores com o pré-processamento na ferramenta Weka.

Amostragem	ME	BIM	J48	SVM
Resample	86,9923	82,8759	84,7969	**89,8463**
SpreadSubsample	68,8942	76,3061	72,0534	75,9416

ME:Máxima Entropia; BIM:Bayesiano Ingênuo Multinomial; SVM: Máquina de Vetor de Suporte.

Fonte: Autoria Própria

optou-se por realizar a etapa de pré-processamento do texto de forma externa a ferramenta Weka. Para a realização disso foi utilizada a ferramenta NLTK, com a qual foi realizada a tokenização do texto, a remoção de *stopwords*, além da análise da diferença do desempenho do classificador com o uso de *stemming*, a aplicação de ambas as técnicas é aprofundada no trabalho de Savoy (1999). Os resultados dessa análise podem ser visualizados na Tabela 2 sem *stemming* e na Tabela 3 com *stemming*.

Tabela 2 – Percentuais do desempenho dos classificadores com o pré-processamento utilizando a ferramenta NLTK.

FM	Atributos	Amostragem	ME	BIM	J48	SVM
1	8603	Resample	84,4127	81,8331	80,4610	**87,3216**
		SpreadSubsample	67,1931	74,1190	69,7448	75,4556
2	3107	Resample	85,7299	79,9670	80,2414	86,2788
		SpreadSubsample	82,8210	74,3620	69,7448	75,2126
3	2032	Resample	82,8210	79,5279	80,1866	85,2360
		SpreadSubsample	57,7156	74,2405	69,7448	75,8201
4	1566	Resample	80,6805	78,5400	80,1866	84,3029
		SpreadSubsample	59,0522	73,8760	69,7448	74,4835
5	1229	Resample	83,0406	78,1009	80,2963	84,6322
		SpreadSubsample	57,7156	73,8760	69,7448	74,6051
6	1047	Resample	85,7848	78,1009	80,2414	82,8759
		SpreadSubsample	59,2952	73,9975	69,7448	73,9975

FM:Frequência Mínima;Atributos:a quantidade de atributos; ME:Máxima Entropia; BIM:Bayesiano Ingênuo Multinomial; SVM: Máquina de Vetor de Suporte.

Fonte: Autoria Própria

Após a análise dos resultados obtidos com o uso do pré-processamento externo, ficou claro que o uso do pré-processamento interno da ferramenta Weka, apresentado na Tabela 1 era a melhor opção para a sequência do trabalho do que as outras abordagens com a ferramenta NLTK que tiveram seus resultados dispostos nas Tabelas 2 e 3, devido a escolha em obter o melhor aproveitamento possível, mesmo que representado uma maior dimensionalidade.

Por fim, para a melhora do desempenho foi utilizado o algoritmo do Gradiente Descendente Estocástico, o qual otimiza o resultado de outros algoritmos para a otimização da

Tabela 3 – Percentuais do desempenho dos classificadores com o pré-processamento utilizando a ferramenta NLTK e aplicação de *stemming*.

FM	Atributos	Amostragem	ME	BIM	J48	SVM
1	5958	Resample	85,4555	80,2414	80,6805	**86,1141**
		SpreadSubsample	65,1275	74,1190	70,8383	71,6889
2	2343	Resample	83,3150	79,0340	80,7903	83,9187
		SpreadSubsample	61,7253	74,8481	70,8383	71,4459
3	1667	Resample	82,1075	78,6498	80,7903	82,9308
		SpreadSubsample	60,0243	74,9696	70,8383	71,9319
4	1333	Resample	82,6015	77,7167	80,7903	82,4917
		SpreadSubsample	60,2673	75,2126	70,8383	71,5674
5	1118	Resample	82,8210	77,0581	80,7903	82,1624
		SpreadSubsample	59,4167	74,1190	70,8383	71,0814
6	949	Resample	82,7661	76,6190	80,1317	81,8880
		SpreadSubsample	60,3888	73,7545	70,8383	71,4459

FM:Frequência Mínima;Atributos:a quantidade de atributos; ME:Máxima Entropia; BIM:Bayesiano Ingênuo Multinomial; SVM: Máquina de Vetor de Suporte.

Fonte: Autoria Própria

divisão em duas classes, o que se adéqua perfeitamente ao objetivo dessa classificação de dividir os *tweets* entre os que indicam casos de dengue e os que não indicam. Um aprofundamento em relação as características do algoritmo, seu funcionamento e aplicação pode ser encontrado no trabalho de Sharma (2018). A ferramenta Weka disponibiliza a possibilidade de utilizar a saída dos algoritmos SVM e Máxima Entropia, o que é interessante tendo em vista que esses foram os métodos com melhor desempenho nas Tabelas 1, 2 e 3. Os resultados podem ser visualizados na Tabela 4, na qual também estão apresentados testes com validação cruzada, aprofundado em Berrar (2019), além dos testes com divisão do conjunto de dados, para verificar de forma mais ampla o desempenho do classificador.

Tabela 4 – Percentuais do desempenho dos classificadores com o método do Gradiente Descendente Estocástico.

Teste	Máxima Entropia	Máquina de Vetor de Suporte
Divisão do Dataset	90,2359	90,1262
Validação Cruzada	91,1600	91,5144

Fonte: Autoria Própria

O desempenho de ambos os algoritmos foi muito próximo, com a maior diferença sendo de 0,35%, sendo essa a vantagem do algoritmo de Máquina de Vetor de Suporte em relação ao algoritmo de Máxima Entropia, o que foi o fator decisivo para a escolha do método de SVM para a criação do modelo classificador.

4.2 CAPTURA DOS TWEETS GEORREFERENCIADOS E CLASSIFICAÇÃO

Inicialmente o objetivo era de capturar *tweets* de todas as regiões do estado do Paraná, entretanto a organização do georreferenciamento do Twitter no Brasil não tem o mesmo padrão estadunidense, que é utilizado de exemplo pela aplicação, não sendo possível especificar exatamente o estado ou região de interesse.

Por isso foram definidas áreas de captura para os *tweets*. Para a definição dessas áreas foram levadas em consideração as regiões do estado com maior incidência de dengue no período de 1 de janeiro de 2016 a 31 de dezembro de 2018, segundo os dados obtidos do sistema InfoDengue, resultando no mapa que pode ser visto na Figura 16. E na Tabela 5, são apresentadas as 10 cidades usadas como alvo para a determinação das regiões de captura.

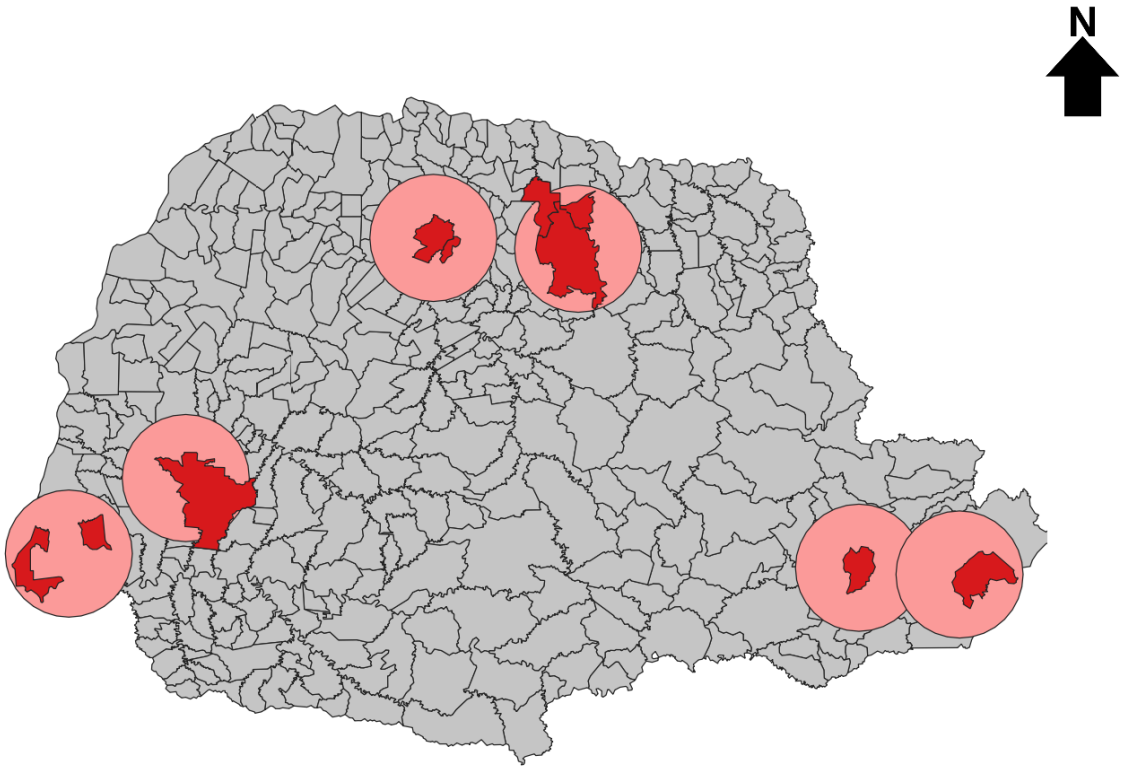

Figura 16 – Regiões de captura de *tweets* no estado do Paraná com as cidades com maior incidência de dengue em destaque.

Fonte: Autoria Própria

Entretanto, com essas capturas não foi possível obter uma quantidade considerável de *tweets*. Por isso, foi utilizada a abordagem de obter a posição georreferenciada de acordo com o que é definido pelo usuário como sua casa. As cidades selecionadas podem ser visualizadas

Tabela 5 – As 10 cidades com maior incidência de casos de dengue entre 2016 e 2018 com os respectivos valores totais de casos no período segundo o sistema InfoDengue.

Cidade	Número total de casos
Paranaguá	29.219
Londrina	28.752
Foz Do Iguaçu	23.221
Maringá	22.752
Curitiba	15.583
Medianeira	7.357
Cascavel	6.957
Cambe	5.934
Ibiporã	4.840
Sarandi	4.583

Fonte: Autoria Própria

no mapa da Figura 17 e na Tabela 6.

Os resultados dessas duas abordagens podem ser visualizados na Tabela 7

Como pode-se ver os *tweets* georreferenciados pela identificação da casa estão presentes em uma quantidade muito maior do que os *tweets* georreferenciados exatamente, e mesmo que proporcionalmente os *tweets* georreferenciados exatamente apresentem uma maior eficiência na identificação de casos de dengue, eles foram desconsiderados para o restante do trabalho em virtude da sua baixa distribuição, já que a correspondência entre os *tweets* identificados pela casa tendem a ter uma eficácia consideravelmente maior.

4.3 AGRUPAMENTO DOS TWEETS COM DENGUE

Para realizar o agrupamento dos *tweets* foi definido que o mínimo de *tweets* a ser considerado seria de 2. Além disso foram definidas 3 opções para a variável de distância que é denominada EPS (épsilon de variação máxima de distância que o algoritmo considera). Essas opções se referem principalmente ao fator se cidades próximas serão consideradas no agrupamento ou não.

Com isso é possível perceber que não existiu diferença entre o uso de um épsilon de distância de 0,05 e 0,10, pois ambos agruparam exatamente o mesmo número de itens: 1184, já quando ambos são comparados ao épsilon de 0,20 existe uma diferença de 6 itens, e por isso a sequência do trabalho irá considerar a comparação entre o épsilon de 0,20 e 0,10, já que o

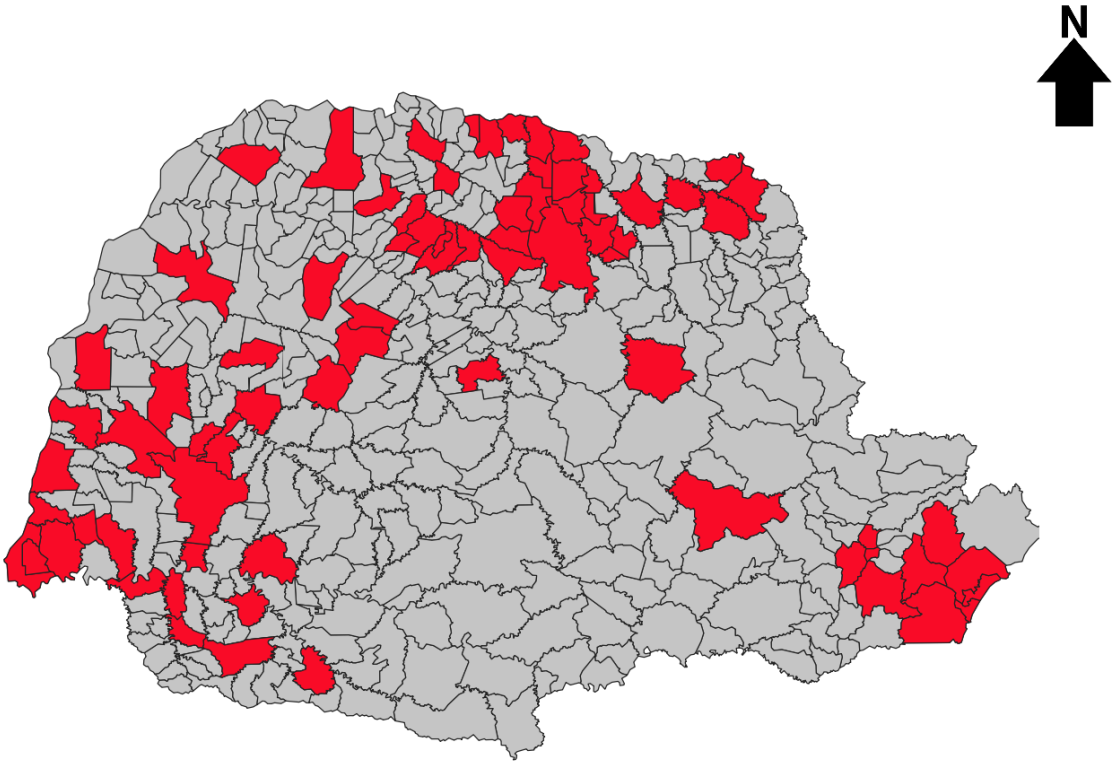

Figura 17 – Cidades de captura de *tweets* no estado do Paraná.

Fonte: Autoria Própria

Tabela 6 – As 76 cidades com maior incidência de casos de dengue no estado entre janeiro de 2016 e dezembro 2018 com os respectivos valores totais de casos no período.

Cidade	Casos	Cidade	Casos
Paranaguá	29.219	Guaratuba	966
Londrina	28.752	Loanda	952
Foz Do Iguaçu	23.221	Assis Chateaubriand	935
Maringá	22.752	Alvorada Do Sul	916
Curitiba	15.583	Capanema	898
Medianeira	7.357	Cambara	862
Cascavel	6.957	Matelândia	827
Cambe	5.934	Campo Mourão	772
Ibiporã	4.840	Bandeirantes	766
Sarandi	4.583	Mamborê	761
Paranavaí	3.570	Mandaguari	735
Santa Terezinha De Itaipu	2.877	Lupionópolis	718
Assai	2.856	Cafelândia	711
Francisco Beltrão	2.347	Centenário Do Sul	699
Umuarama	2.264	Jacarezinho	699
Rolândia	2.221	Pontal Do Parana	688
Santa Helena	2.116	Santa Cecilia Do Pavão	684
Toledo	2.070	Itaipulândia	680
São Miguel Do Iguaçu	1.950	Rancho Alegre	677
Colorado	1.902	Marechal Cândido Rondon	666
Ampere	1.882	São Sebastião Da Amoreira	664
Cianorte	1.687	Morretes	646
Jataizinho	1.642	Matinhos	639
Apucarana	1.596	Porecatu	635
Paiçandu	1.533	Pinhais	627
Arapongas	1.434	Peabiru	598
Sertanópolis	1.420	Santa Fé	586
Marialva	1.307	Realeza	570
Bela Vista Do Paraíso	1.196	Mandaguaçu	567
Ubiratã	1.117	Colombo	556
Cornélio Procópio	1.101	Santo Antônio Da Platina	532
São José Dos Pinhais	1.079	Ponta Grossa	527
Pato Branco	1.069	Nova Esperança	524
Antonina	1.067	Telêmaco Borba	521
Corbélia	1.057	Boa Vista Da Aparecida	520
Quedas Do Iguaçu	1.007	Dois Vizinhos	517
Goioerê	999	Terra Roxa	516
Ivaiporã	967	Primeiro De Maio	496

Fonte: Autoria Própria

Tabela 7 – Resultado da classificação dos *tweets*.

Origem	Total	Identificados com dengue
Georreferenciados exatamente	82	14
Georreferenciados pela casa	10380	1485
Total	10412	1567

Fonte: Autoria Própria

Tabela 8 – Resultado da quantidade de itens agrupados na distribuição semanal dos itens de acordo com a distribuição semanal.

EPS	Distância em KM	Itens agrupados
0,05	5,6	1184
0,10	11,1	1184
0,20	22,2	1190

EPS: Épsilon de variação máxima de distância que o algoritmo considera.

Fonte: Autoria Própria

épsilon de 0,10 representará também o épsilon de 0,05.

4.4 COMPARAÇÃO ENTRE CASOS REAIS E OS CASOS IDENTIFICADOS PELO TWITTER

Para realizar a comparação entre os resultados obtidos por meio da identificação pelos *tweets* e os dados referentes aos casos reais, foi definada uma métrica em que foram selecionadas as cidades com maior incidência de dengue na semana de acordo com um limite, que poderia ser 5, 10 ou 20 cidades. Cada uma dessas listas de cidades então foi comparada com a lista de cidades indicada pela identificação de *tweets*, resultando em 3 avaliadores: quantidade correta, que se refere a quantos itens estiveram presentes tanto na identificação de *tweets* quanto na de casos reais; quantidade faltante, que se refere a quantos itens estavam na lista de casos reais mas não na lista de casos identificados pelos *tweets*; e quantidade excedente, que se refere a quantos itens estavam presentes na lista de casos identificados pelos *tweets* mas não na lista de casos reais. Os resultados dessa comparação podem ser visualizados na Tabela 9.

Com a análise da tabela, é possível constatar que o melhor desempenho de acerto foi o obtido com limites menores de cidade, enquanto o melhor desempenho de não erro foi obtido com um número maior de cidades, sendo o épsilon de 0,20 com o limite de cidades de 5 o

Tabela 9 – Resultados obtidos com a comparação entre as listas de cidades com maior incidência de casos reais na semana e as cidades identificadas por meio dos casos identificados pelo Twitter na semana.

EPS	Limite	Quantidade correta	Quantidade faltante	Quantidade excedente
0,10	5	178	572	32
	10	203	1297	38
	20	207	2793	7
0,20	5	178	572	13
	10	203	1297	3
	20	208	2792	8

EPS: Épsilon de variação máxima de distância que o algoritmo considera; Limite: quantidade de cidades da lista de cidades com maior incidência ordenada pela incidência no período; Quantidade correta: o valor total de itens que estão presentes tanto na lista de cidades com maior incidência e na lista identificada por meio dos *tweets*; Quantidade faltante: o valor total de itens que está presente na lista de cidades com maior incidência mas está ausente na lista de cidades identificadas pelos *tweets*; Quantidade excedente: o valor total de itens que estão presentes na lista de cidades identificadas pelos *tweets* mas não na lista de cidades com maior incidência.

Fonte: Autoria Própria

caso que mantém melhor acerto sem um valor tão elevado de erro, pois esse desempenho obtido representa que a abordagem foi capaz de identificar 23,73% das cidades com maior incidência de dengue, apresentando um erro de apenas 6,81%.

Com a percepção apresentada nesses resultados de que o maior índice de acerto se encontrava com um limite menor de cidades, decidiu-se averiguar qual era o desempenho quando o limite para a quantidade de cidades com maior incidência tendo como base os casos reais fosse ainda menor, o resultado dessa comparação pode ser visto na Tabela 10.

Tabela 10 – Resultados obtidos com a comparação entre as listas de cidades com maior incidência de casos reais na semana e as cidades identificadas por meio dos casos identificados pelo Twitter na semana com um limite menor do que 5.

EPS	Limite	Quantidade correta	Quantidade faltante	Quantidade excedente
0,10	2	74	226	136
	3	111	339	99
	4	143	457	67
0,20	2	74	226	142
	3	111	339	105
	4	143	457	73

EPS: Épsilon de variação máxima de distância que o algoritmo considera; Limite: quantidade de cidades da lista de cidades com maior incidência ordenada pela incidência no período; Quantidade correta: o valor total de itens que estão presentes tanto na lista de cidades com maior incidência e na lista identificada por meio dos *tweets*; Quantidade faltante: o valor total de itens que está presente na lista de cidades com maior incidência mas está ausente na lista de cidades identificadas pelos *tweets*; Quantidade excedente: o valor total de itens que estão presentes na lista de cidades identificadas pelos *tweets* mas não na lista de cidades com maior incidência.

Fonte: Autoria Própria

Com a análise feita desses resultados obtidos, é possível constatar que os melhores resultados em questão de taxa de acerto foram com um menor limite de cidades, com o valor de aproveitamento sendo de 24,67% tanto para o limite de 2 e 3 e 23,83% para o limite de

4 cidades. Enquanto da mesma forma que anteriormente o menor incide de erros se encontra com os maiores limites, sendo o menor deles de 31,91% obtido com épsilon 0,10 e limite de 4 cidades.

Por fim foi feita uma análise considerando um diferente nível de granularidade, para isso as semanas foram agrupadas em 4, para que se tivesse algo próximo a uma análise mensal, os resultados dessa análise estão dispostos na Tabela 11.

Tabela 11 – Resultados obtidos com a comparação entre as listas de cidades com maior incidência de casos reais na semana e as cidades identificadas por meio dos casos identificados pelo Twitter na semana com uma granularidade temporal próxima de mensal.

EPS	Limite	Quantidade correta	Quantidade faltante	Quantidade excedente
0,10	2	27	36	90
	3	32	52	85
	4	38	75	79
	5	48	89	69
	10	71	185	46
	20	104	413	13
0,20	2	27	36	96
	3	32	52	91
	4	38	75	85
	5	48	89	75
	10	71	185	52
	20	104	413	19

EPS: Épsilon de variação máxima de distância que o algoritmo considera; Limite: quantidade de cidades da lista de cidades com maior incidência ordenada pela incidência no período; Quantidade correta: o valor total de itens que estão presentes tanto na lista de cidades com maior incidência e na lista identificada por meio dos *tweets*; Quantidade faltante: o valor total de itens que está presente na lista de cidades com maior incidência mas está ausente na lista de cidades identificadas pelos *tweets*; Quantidade excedente: o valor total de itens que estão presentes na lista de cidades identificadas pelos *tweets* mas não na lista de cidades com maior incidência.

Fonte: Autoria Própria

Com essa análise em uma granularidade diferente, foi possível obter resultados de acerto em até 42,85% e de erro em até 11,11%, fica evidente como uma maior quantidade de dados pode tornar os resultados melhores, diminuindo a importância de casos discrepantes e aproximando os resultados obtidos por meio de *tweets* aos casos reais.

5 DISCUSSÕES

Nesse capítulo serão discutidos os resultados obtidos no capítulo anterior, abordando algumas vantagens, desvantagens e buscando expandir possibilidades.

Como primeiro ponto, é preciso abordar as diferentes metodologias de pré-processamento, por meio da análise das Tabelas 1, 2 e 3 é possível fazer a observação de que a melhor das abordagens analisadas para o pré-processamento para texto é a implementada nativamente na ferramenta Weka.

Após a análise das mesmas tabelas, ficou evidente o melhor aproveitamento dos algoritmos de Máxima Entropia e Máquina de Vetor de Suporte, além da amostragem Resample, que posteriormente foram analisados na Tabela 4, fazendo uso do algoritmo de Gradiente Descendente Estocástico. Isto proporcionou sempre resultados superiores a 90%, sendo a escolha pelo algoritmo de Máquina de Vetor de Suporte justificável por ter obtido uma pequena margem de melhor aproveitamento da mesma.

Na sequência, com a obtenção dos *tweets* georreferenciados, ficou evidente por meio dos valores mostrados na Tabela 7 que a obtenção dos *tweets* pela identificação da casa do usuário, com uma proporção 12.658% maior do que os *tweets* georreferenciados exatamente.

Já com a análise dos dados das Tabelas 9 e 10 é perceptível que quanto menor a quantidade de cidades selecionadas para a comparação, maior seria o aproveitamento entre a quantidade correta e a quantidade faltante, ao mesmo tempo em que quanto maior a quantidade de cidades maior o aproveitamento entre a quantidade correta e a quantidade excedente, sendo o valor de 5 cidades o que proporciona o maior equilíbrio.

Entretanto, o percentual de acerto ainda foi considerado abaixo do potencial esperado, e por isso optou-se por verificar de forma separa a análise espacial e a análise temporal. Para a análise temporal foram produzidos gráficos da distribuição temporal dos casos de dengue e dos *tweets* identificados com dengue. Para uma melhor apresentação do conteúdo foi desenvolvido um gráfico apresentado na Figura 18, onde pode ser visualizado ao mesmo tempo a quantidade de casos por semana e a quantidade de *tweets* por semana, de maneira que os dados não tenham sua visualização prejudicada pela diferença de mais de 150 vezes entre as quantidades médias, além de permitir a visualização dos dados quantitativos para cada respectivo valor.

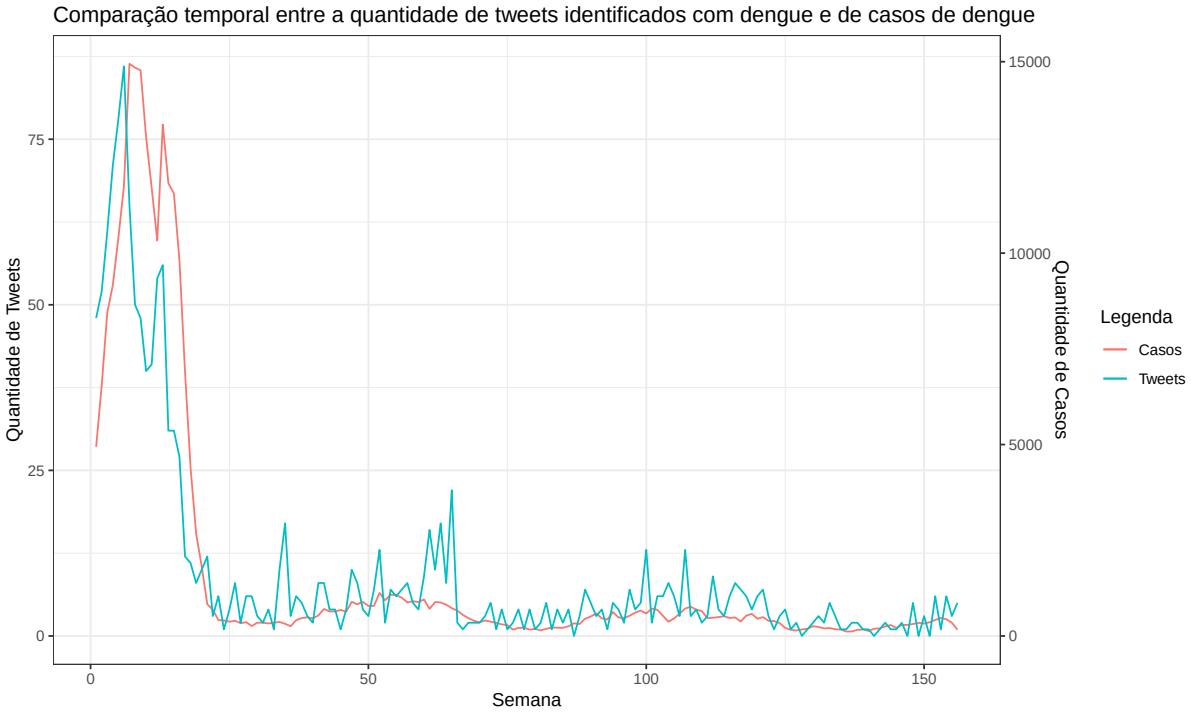

Figura 18 – Distribuição temporal da quantidade dos *tweets* identificados com dengue e dos casos de dengue pelas semanas epidemiológicas do período entre 2016 e 2018.

Fonte: Autoria Própria

Com a análise dos gráficos é possível visualizar que apesar da grande diferença proporcional entre os dados existe uma similaridade entre os dados, com a quantidade de *tweets* identificados com dengue possuindo uma variação mais brusca do que os casos reais.

Já quanto a distribuição espacial a análise foi feita por meio de uma tabela de comparação como a utilizada anteriormente, mas que dessa vez desconsiderou a questão das semanas epidemiológicas, utilizando listas com as 10, 20 e 30 cidades com maior incidência de dengue no período entre janeiro de 2016 e dezembro de 2018 e comparando com os *tweets* identificados com dengue do mesmo período, o resultado disso pode ser visto na Tabela 12.

Com a visualização da tabela é possível perceber como os aproveitamentos de acertos (entre 46,67% e 80%), ausências (entre 20% e 53,34%) e erros (entre 7,70% e 47,37%), apresentando resultados consideravelmente melhores dos que obtidos quando o fator temporal foi considerado, especialmente em listas com 20 cidades que apresentam bons resultados em ambos os fatores de comparação.

Para proporcionar uma visualização mas clara disso foram desenvolvidos mapas de calor da distribuição de casos de dengue e dos *tweets* identificados com dengue entre 2016 e 2018, ambos podem ser visualizados respectivamente nas imagens 19 e 20.

Outra análise ainda feita foi sobre a relação entre *tweets* identificados com dengue e

Tabela 12 – Resultados obtidos com a comparação entre as listas de cidades com maior incidência de casos reais nos três anos e as cidades identificadas por meio dos casos identificados pelo Twitter nos três anos.

EPS	Limite	Quantidade correta	Quantidade faltante	Quantidade excedente
0,10	10	8	2	7
	20	12	8	3
	30	14	16	5
0,20	10	8	2	9
	20	12	8	1
	30	14	16	3

EPS: Épsilon de variação máxima de distância que o algoritmo considera; Limite: quantidade de cidades da lista de cidades com maior incidência ordenada pela incidência no período; Quantidade correta: o valor total de itens que estão presentes tanto na lista de cidades com maior incidência e na lista identificada por meio dos *tweets*; Quantidade faltante: o valor total de itens que está presente na lista de cidades com maior incidência mas está ausente na lista de cidades identificadas pelos *tweets*; Quantidade excedente: o valor total de itens que estão presentes na lista de cidades identificadas pelos *tweets* mas não na lista de cidades com maior incidência.

Fonte: Autoria Própria

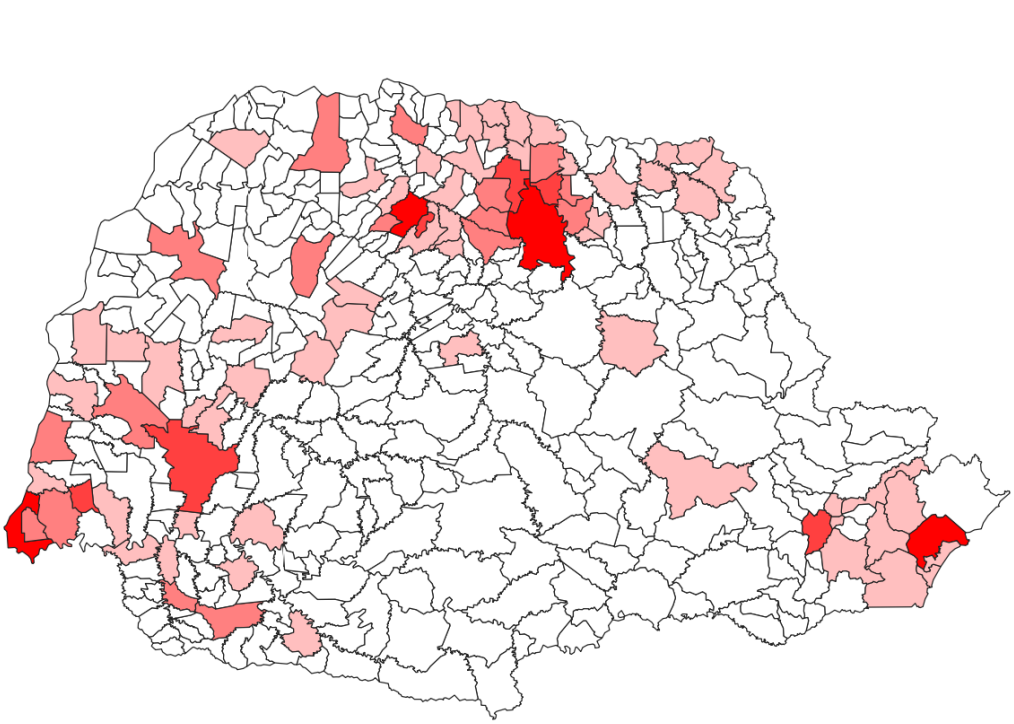

Figura 19 – Mapa de calor dos casos de dengue no Paraná no período entre 2016 e 2018.

Fonte: Autoria Própria

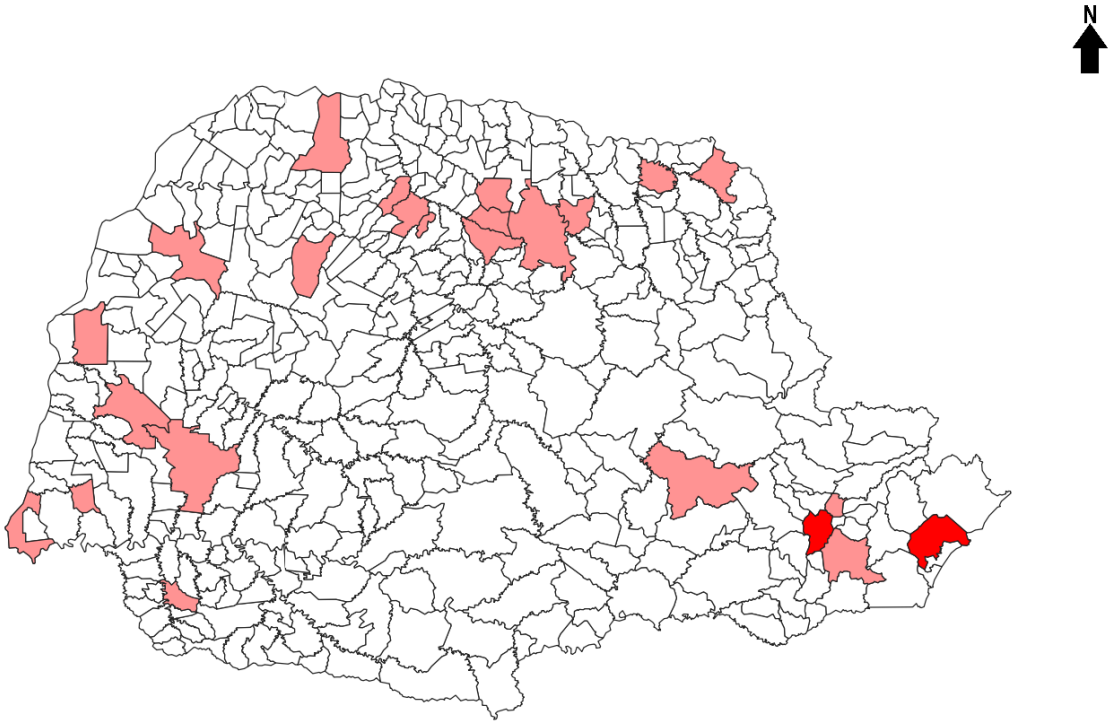

Figura 20 – Mapa de calor dos *tweets* identificados com dengue nos Paraná no período entre 2016 e 2018.

Fonte: Autoria Própria

cidades com maior incidência de dengue no período comparado com a relação entre os *tweets* e a população das cidades. Para isso foi construído o gráfico apresentado na Figura 21, onde as quantidades de *tweets* por cidade são apresentados ordenados pela população das cidades (da maior para a menor população), e ordenados pela incidência de dengue (da cidade com maior incidência para a com a menor).

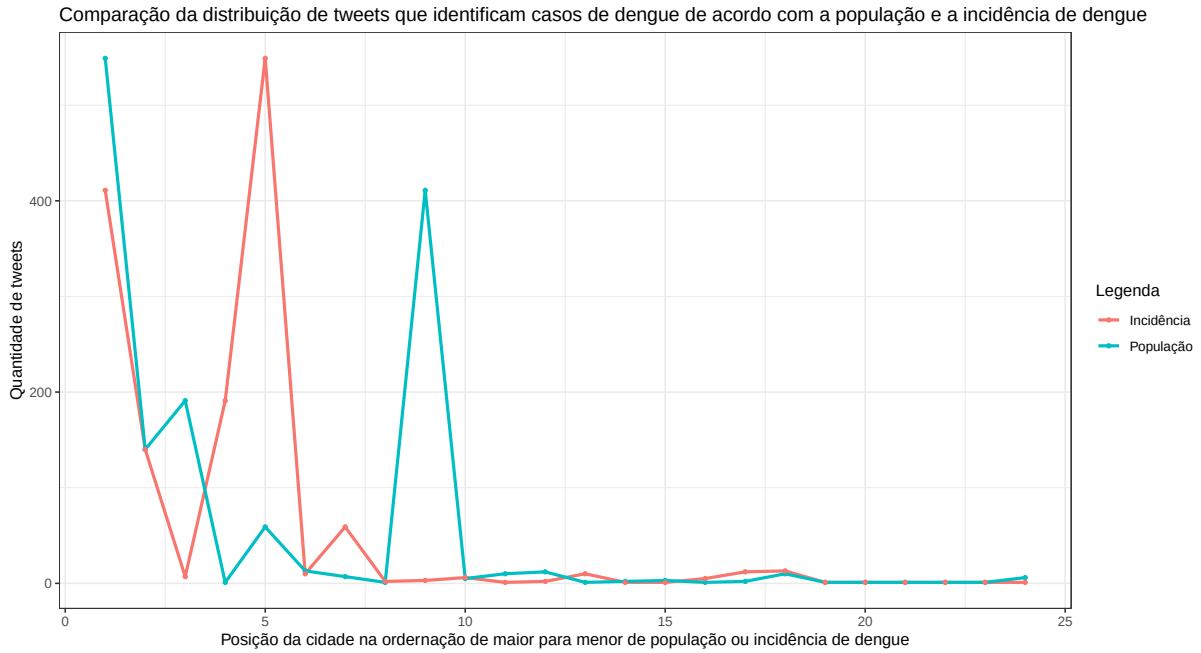

Figura 21 – Distribuição *tweets* identificados com dengue de acordo com a população das cidades e a incidência de dengue das cidades.

Fonte: Autoria Própria

Nesse gráfico é possível visualizar que tanto a população das cidades quanto a incidência de dengue são fatores que interferem na quantidade de *tweets* que está presente, mesmo em um trabalho onde somente o segundo fator foi o ponto de interesse para a busca, sendo esse um fator que deve ser levado em consideração, podendo inclusive ser estabelecida uma relação mais forte entre os dois fatores, o que carece de uma análise, especialmente com o uso de uma amostragem mais ampla.

Com tanto o fator espacial quanto o temporal analisados separadamente, fica evidente que o uso de *tweets* que sejam identificados com dengue possui um potencial considerável de identificar casos de dengue em ambos os fatores, sendo o desempenho na análise espaço-temporal com possibilidade de melhora em caso de um volume maior de dados.

6 CONSIDERAÇÕES FINAIS

6.1 CONCLUSÕES

A execução deste trabalho possibilitou verificar que o potencial de identificação espaço-temporal de casos de dengue por meio do uso de *tweets* identificados com casos diretos de dengue, desenvolvendo um modelo classificador para os *tweets* e uma abordagem comparativa direta com casos reais.

Na elaboração do modelo classificador pode-se concluir que o uso do pré-processamento interno da ferramenta Weka, com o uso da amostragem Resample foi a abordagem mais vantajosa, e que em relação à abordagem de classificação o uso do algoritmo de Máquina de Vetor de Suporte de forma conjunta com o Gradiente Descendente Estocástico foi a mais vantajosa.

Já para a abordagem de obtenção de *tweets* georreferenciados concluiu-se que a obtenção pela identificação do que o usuário indica como sua casa é mais interessante do que a abordagem em que os *tweets* são obtidos pelo georreferenciamento direto.

Por fim foi possível concluir que tanto a análise espacial quanto a temporal são aplicáveis, enquanto a análise espaço-temporal apesar de ter resultados interessantes, apresenta um resultado abaixo de suas análises isoladas, o que pode ser atribuído principalmente a uma menor quantidade de dados quando a distribuição considera os dois fatores.

6.2 TRABALHOS FUTUROS

No decorrer do trabalho fatores de limitação de tempo e recursos financeiros impediram que algumas análises pudessem ser feitas, sendo essas análises divididas de acordo com o fluxo de execução do trabalho.

Na etapa de escolha do classificador três fatores podem ser destacados, sendo o primeiro deles a possibilidade do uso de uma quantidade maior de *tweets* rotulados para o treinamento do classificador. Já a escolha do pré-processamento é uma etapa que permite a exploração de uma gama maior de alternativas, da mesma maneira que com as opções de classificador é possível realizar um estudo mais amplo de alternativas caso um período maior de tempo esteja disponível.

Para a obtenção dos *tweets* georreferenciados a análise de mais cidades, com uma expansão para outros estados e em um período maior de tempo pode permitir uma análise mais aprofundada, que está ligada à disponibilidade de recursos financeiros para um maior acesso à API do Twitter.

Por fim, como métodos de agrupamento é possível realizar uma comparação entre diferentes agrupamentos por densidade, de forma a entender as diferenças que cada alternativa pode causar na comparação final, sendo esse novamente um estudo dependente de um período maior disponível.

REFERÊNCIAS

AALST, W. M. Van der. **Process mining: data science in action**. Berlim (Alemanha): Springer, 2016.

AGGARWAL, C. C. **Data classification: algorithms and applications**. Boca Raton (Estados Unidos da América): CRC press, 2014.

AGGARWAL, C. C. **Data mining: the textbook**. Berlim (Alemanha): Springer, 2015.

AGGARWAL, C. C.; ZHAI, C. X. **Mining Text Data**. Berlim (Alemanha): Springer Publishing Company, Incorporated, 2012.

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Database mining: A performance perspective. **IEEE transactions on knowledge and data engineering**, IEEE, v. 5, n. 6, p. 914–925, 1993.

ALBUQUERQUE, J. P. D. et al. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. **International Journal of Geographical Information Science**, Taylor & Francis, v. 29, n. 4, p. 667–689, 2015.

ALLEN, C. et al. Applying gis and machine learning methods to twitter data for multiscale surveillance of influenza. **PloS one**, Public Library of Science, v. 11, n. 7, p. e0157734, 2016.

ALUÍSIO, S.; CUNHA, A.; SCARTON, C. Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In: **International Conference on Computational Processing of the Portuguese Language**. Tomar (Tomar): [s.n.], 2016. p. 109–114.

BATHRELLOS, G. D. et al. Suitability estimation for urban development using multi-hazard assessment map. **Science of the Total Environment**, Elsevier, v. 575, p. 119–134, 2017.

BENEDETTO, F.; TEDESCHI, A. Big data sentiment analysis for brand monitoring in social media streams by cloud computing. In: PEDRYCZ, W.; CHEN, S.-M. (Ed.). **Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence**. [S.l.: s.n.], 2016.

BERRAR, D. Cross-validation. In: RANGANATHAN, S. et al. (Ed.). **Encyclopedia of Bioinformatics and Computational Biology**. Oxford (Reino Unido): Academic Press, 2019. p. 542 – 545.

BHATT, S. et al. The global distribution and burden of dengue. **Nature**, Nature Publishing Group, v. 496, n. 7446, p. 504–507, 2013.

BIVAND, R. S. et al. **Applied spatial data analysis with R**. Berlim (Alemanha): Springer, 2008.

BOULOS, M. N. K. et al. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, ogc standards and application examples. **International journal of health geographics**, BioMed Central, v. 10, n. 1, p. 67, 2011.

BRAMER, M. **Principles of data mining**. Berlim (Alemanha): Springer, 2016.

BRASIL, M. d. S. Boletim epidemiológico de n 46. **Secretaria de Vigilância em Saúde Ministério da Saúde**, v. 46, 2015.

BRASIL, M. d. S. Boletim epidemiológico de n 38. **Secretaria de Vigilância em Saúde Ministério da Saúde**, v. 47, 2016.

BRASIL, M. d. S. Boletim epidemiológico de n 45. **Secretaria de Vigilância em Saúde Ministério da Saúde**, v. 48, 2017.

CÂMARA, G. et al. Spring: Integrating remote sensing and gis by object-oriented data modelling. **Computers & graphics**, Elsevier, v. 20, n. 3, p. 395–403, 1996.

CHEN, M.-S.; HAN, J.; YU, P. S. Data mining: an overview from a database perspective. **IEEE Transactions on Knowledge and data Engineering**, IEEE, v. 8, n. 6, p. 866–883, 1996.

CHENG, Z.; CAVERLEE, J.; LEE, K. You are where you tweet: a content-based approach to geo-locating twitter users. In: ACM. **Proceedings of the 19th ACM international conference on Information and knowledge management**. Nova Iorque (Estados Unidos da América), 2010. p. 759–768.

CIOS, K. J. et al. **Data mining: a knowledge discovery approach**. Berlim (Alemanha): Springer Science & Business Media, 2007.

CODECO, C. et al. InfoDengue: a nowcasting system for the surveillance of dengue fever transmission. **bioRxiv**, Cold Spring Harbor Laboratory, 2016.

CORTI, P. et al. **PostGIS Cookbook**. Birmingham (Reino Unido): Packt Publishing Ltd, 2014.

CRAMPTON, J. W. **Mapping: A critical introduction to cartography and GIS**. Trenton (Estados Unidos da América): John Wiley & Sons, 2011.

DALE, P. **Mathematical techniques in GIS**. Boca Raton (Estados Unidos da América): CRC Press, 2014.

DIXON, B.; UDDAMERI, V.; RAY, C. **GIS and Geocomputation for Water Resource Science and Engineering**. Trenton (Estados Unidos da América): John Wiley & Sons, 2015.

DREDZE, M. et al. Carmen: A twitter geolocation system with applications to public health. In: **Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence**. Bellevue (Estados Unidos da América): [s.n.], 2013.

FARHADLOO, M.; ROLLAND, E. Fundamentals of sentiment analysis and its applications. In: PEDRYCZ, W.; CHEN, S.-M. (Ed.). **Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence**. [S.l.: s.n.], 2016.

FAYYAD, U. Knowledge discovery in databases: An overview. In: LAVRAČ, N.; DŽEROSKI, S. (Ed.). **Inductive Logic Programming**. Berlim (Alemanha): Springer Berlin Heidelberg, 1997. p. 1–16.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

FEI, G. et al. A dictionary-based approach to identifying aspects implied by adjectives for opinion mining. **Proceedings of COLING 2012: Posters**, p. 309–318, 2012.

FERNANDO, E.; JESUS, R. D. Interface entre a Climatologia e a Epidemiologia: uma abordagem geográfica. **GeoTextos**, v. 6, p. 211–236, 2010.

GHOSH, D.; GUHA, R. What are we 'tweeting'about obesity? mapping tweets with topic modeling and geographic information system. **Cartography and geographic information science**, Taylor & Francis, v. 40, n. 2, p. 90–102, 2013.

GIRI, C. et al. Status and distribution of mangrove forests of the world using earth observation satellite data. **Global Ecology and Biogeography**, Wiley Online Library, v. 20, n. 1, p. 154–159, 2011.

GOMIDE, J. et al. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In: **Proceedings of the 3rd International Web Science Conference**. Nova Iorque (Estados Unidos da América): ACM, 2011. (WebSci '11), p. 3:1–3:8.

GUAGLIARDO, M. F. Spatial accessibility of primary care: concepts, methods and challenges. **International journal of health geographics**, BioMed Central, v. 3, n. 1, p. 3, 2004.

GUZELLA, T. S.; CAMINHAS, W. M. A review of machine learning approaches to spam filtering. **Expert Systems with Applications**, v. 36, n. 7, p. 10206 – 10222, 2009. ISSN 0957-4174.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. São Francisco (Estados Unidos da América): [s.n.], 2012.

HASTIE ROBERT TIBSHIRANI, J. F. T. **The elements of statistical learning: Data mining, inference, and prediction**. 2nd ed.. ed. Berlim (Alemanha): Springer, 2009. (Springer Series in Statistics).

HEYWOOD, D. I.; CORNELIUS, S. C.; CARVER, S. **An introduction to geographical information systems**. Upper Saddle River (Estados Unidos da América): Pearson Prentice Hall, 2011.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: ACM. **Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2004. p. 168–177.

HUANG, Y.; SHEKHAR, S.; XIONG, H. Discovering colocation patterns from spatial data sets: a general approach. **IEEE transactions on knowledge & data engineering**, IEEE, n. 12, p. 1472–1485, 2004.

JI, X. et al. Twitter sentiment classification for measuring public health concerns. **Social Network Analysis and Mining**, Springer, v. 5, n. 1, p. 13, 2015.

KIMBALL, R.; ROSS, M. **The data warehouse toolkit: the complete guide to dimensional modeling**. Trenton (Estados Unidos da América): John Wiley & Sons, 2011.

KOPERSKI, K.; HAN, J. Discovery of spatial association rules in geographic information databases. In: **International Symposium on Spatial Databases**. [S.l.: s.n.], 1995. p. 47–66.

KUHN, M.; JOHNSON, K. **Applied predictive modeling**. Berlim (Alemanha): Springer, 2013.

KWAK, H. et al. What is twitter, a social network or a news media? In: **Proceedings of the 19th International Conference on World Wide Web**. Nova Iorque (Estados Unidos da América): ACM, 2010. (WWW '10), p. 591–600.

LAROSE, D. T.; LAROSE, C. D. **Discovering knowledge in data: an introduction to data mining**. Trenton (Estados Unidos da América): John Wiley & Sons, 2014.

LAURINI, R. 9 - geographic knowledge discovery and data mining. In: LAURINI, R. (Ed.). **Geographic Knowledge Infrastructure**. [S.l.]: Elsevier, 2017. p. 183 – 194.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. **Mining of massive datasets**. Nova Iorque (Estados Unidos da América): [s.n.], 2014.

LI, D.; WANG, S.; LI, D. **Spatial data mining**. Berlim (Alemanha): Springer, 2015.

LIU, B. **Sentiment analysis: Mining opinions, sentiments, and emotions**. Cambridge (Reino Unido): [s.n.], 2015.

LONGLEY, P. A.; ADNAN, M.; LANSLEY, G. The geotemporal demographics of twitter usage. **Environment and Planning A**, SAGE Publications Sage UK: London, England, v. 47, n. 2, p. 465–484, 2015.

LONGLEY, P. A. et al. **Geographic information systems and science**. Trenton (Estados Unidos da América): John Wiley & Sons, 2005.

LONGUEVILLE, B. D.; SMITH, R. S.; LURASCHI, G. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In: ACM. **Proceedings of the 2009 international workshop on location based social networks**. Nova Iorque (Estados Unidos da América), 2009. p. 73–80.

MA, X. et al. Mining smart card data for transit riders' travel patterns. **Transportation Research Part C: Emerging Technologies**, Elsevier, v. 36, p. 1–12, 2013.

MEIRELLES, F. Pesquisa anual do uso de ti: Administração de recursos de informática. 29ª edição. **GVcia, São Paulo: FGV-EAESP**, 2018.

MINER, G. A. **Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications**. Amsterdã (Países Baixos): [s.n.], 2012.

MUNINGER, M.-I.; HAMMEDI, W.; MAHR, D. The value of social media for innovation: A capability perspective. **Journal of Business Research**, Elsevier, v. 95, p. 116–127, 2019.

O'SULLIVAN, D.; UNWIN, D. **Geographic information analysis**. Trenton (Estados Unidos da América): John Wiley & Sons, 2014.

PACHECO, F. et al. A statistical comparison of neuroclassifiers and feature selection methods for gearbox fault diagnosis under realistic conditions. **Neurocomputing**, Elsevier, v. 194, p. 192–206, 2016.

POZZI, F. A. et al. **Sentiment analysis in social networks**. Boston (Estados Unidos da América): Morgan Kaufmann, 2016.

RATNER, B. **Statistical and Machine-Learning Data Mining, Third Edition: Techniques for Better Predictive Modeling and Analysis of Big Data, Third Edition**. Boca Raton (Estados Unidos da América): CRC Press, 2017.

SAVOY, J. A stemming procedure and stopword list for general french corpora. **J. Am. Soc. Inf. Sci.**, John Wiley & Sons, Inc., v. 50, n. 10, p. 944–952, 1999. ISSN 0002-8231.

SHARMA, A. Guided stochastic gradient descent algorithm for inconsistent datasets. **Applied Soft Computing**, v. 73, p. 1068 – 1080, 2018.

SHELTON, T.; POORTHUIS, A.; ZOOK, M. Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. **Landscape and urban planning**, Elsevier, v. 142, p. 198–211, 2015.

SHEN, J. et al. Real-time superpixel segmentation by dbscan clustering algorithm. **IEEE Transactions on Image Processing**, IEEE, v. 25, n. 12, p. 5933–5942, 2016.

SOUSA, L. et al. VazaDengue: An information system for preventing and combating mosquito-borne diseases with social networks. **Information Systems**, Elsevier Ltd, v. 75, p. 26–42, 2018.

STEIGER, E.; ALBUQUERQUE, J. P. D.; ZIPF, A. An advanced systematic literature review on spatiotemporal analyses of t witter data. **Transactions in GIS**, Wiley Online Library, v. 19, n. 6, p. 809–834, 2015.

STOCK, K. Mining location from social media: A systematic review. **Computers, Environment and Urban Systems**, Elsevier, v. 71, n. March, p. 209–240, 2018.

SUI, D.; GOODCHILD, M. The convergence of gis and social media: challenges for giscience. **International Journal of Geographical Information Science**, Taylor & Francis, v. 25, n. 11, p. 1737–1748, 2011.

SUI, D.; GOODCHILD, M. The convergence of gis and social media: challenges for giscience. **International Journal of Geographical Information Science**, v. 25, n. 11, p. 1737 – 1748, 2011.

TABOADA, M. et al. Lexicon-Based Methods for Sentiment Analysis. **Computational Linguistics**, v. 37, n. 2, p. 267–307, 2011.

THEODORIDIS, S.; KOUTROUMBAS, C. **Pattern Recognition**. Amsterdã (Países Baixos): [s.n.], 2009.

TIAN, B. **GIS technology applications in environmental and earth sciences**. Boca Raton (Estados Unidos da América): CRC Press, 2016.

TOEPKE, S. Structure occupancy curve generation using geospatially enabled social media data. In: . Roma (Itália): [s.n.], 2016. p. 32–38.

VERCELLIS, C. **Business Intelligence: Data Mining and Optimization for Decision Making**. Trenton (Estados Unidos da América): [s.n.], 2009.

WEISS, S. M. et al. **Text mining: predictive methods for analyzing unstructured information**. Berlim (Alemanha): [s.n.], 2005.

WEITZEL, L.; PRATI, R. C.; AGUIAR, R. F. The comprehension of figurative language: What is the influence of irony and sarcasm on nlp techniques? In: PEDRYCZ, W.; CHEN, S.-M. (Ed.). **Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence**. [S.l.: s.n.], 2016.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3rd. ed. São Francisco (Estados Unidos da América): Morgan Kaufmann Publishers Inc., 2011.

WITTEN, I. H. et al. **Data Mining: Practical machine learning tools and techniques**. São Francisco (Estados Unidos da América): Morgan Kaufmann, 2016.

XIAO, J. et al. Evaluating urban expansion and land use change in shijiazhuang, china, by using gis and remote sensing. **Landscape and urban planning**, Elsevier, v. 75, n. 1-2, p. 69–80, 2006.

YANG, S. Y.; MO, S. Y. K. Social media and news sentiment analysis for advanced investment strategies. In: PEDRYCZ, W.; CHEN, S.-M. (Ed.). **Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence**. [S.l.: s.n.], 2016.

YANG, W.; MU, L. Gis analysis of depression among twitter users. **Applied Geography**, Elsevier, v. 60, p. 217–223, 2015.

YOO, S.; SONG, J.; JEONG, O. Social media contents based sentiment analysis and prediction system. **Expert Systems with Applications**, Elsevier Ltd, v. 105, p. 102–111, 2018.

ZHANG, N.; WANG, M.; WANG, N. Precision agriculture—a worldwide overview. **Computers and electronics in agriculture**, Elsevier, v. 36, n. 2-3, p. 113–132, 2002.

ZHENG, Y. Trajectory data mining: an overview. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM, v. 6, n. 3, p. 29, 2015.

ANEXO A – EXEMPLO DE REPRESENTAÇÃO DE TWEET


```
{
        "created_at": "",
        "id_str": "",
        "text": "",
        "user": {
                "id": ,
                "name": "",
                "screen_name": "",
                "location": "",
                "url": "",
                "description": ""
        },
        "place": { },
        "entities": {
                "hashtags": [
                ],
                "urls": [
                        {
                                "url": "",
                                "unwound": {
                                        "url": "",
                                        "title": ""
                                }
                        }
                ],
                "user_mentions": [  ]
        }
}
```

ANEXO B – EXEMPLO DE REPRESENTAÇÃO DE INFORMAÇÃO GEOGRÁFICA DE TWEET


```
{
        "geo": {
        "type": "Point",
        "coordinates": [
                40.74118764,
                -73.9998279
        ]
        },
        "coordinates": {
                "type": "Point",
                "coordinates": [
                        -73.9998279,
                        40.74118764
                ]
        },
        "place": {
                "id": "01a9a39529b27f36",
                "url": "",
                "place_type": "city",
                "name": "Manhattan",
                "full_name": "Manhattan, NY",
                "country_code": "US",
                "country": "United States",
                "bounding_box": {
                        "type": "Polygon",
                        "coordinates": [
                                [
                                        [
```

```
                    -74.026675,
                    40.683935
                ],
                [
                    -74.026675,
                    40.877483
                ],
                [
                    -73.910408,
                    40.877483
                ],
                [
                    -73.910408,
                    40.683935
                ]
            ]
        ]
    },
    "attributes": {

    }
  }
}
```