UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Explorando uma abordagem de Fusão Multimodal para auxílio a Gestão de Desastres: um estudo de caso com tweets e dados contextuais

Thiago Aparecido Gonçalves da Costa

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Thiago Aparecido Gonçalves da Costa

Explorando uma abordagem de Fusão Multimodal para auxílio a Gestão de Desastres: um estudo de caso com tweets e dados contextuais

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Jó Ueyama

USP – São Carlos
Outubro de 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Costa, Thiago Aparecido Gonçalves da
C837 / Thiago Aparecido Gonçalves da Costa;
orientador Jó Ueyama. -- São Carlos, 2021.
193 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Fusão Multimodal. 2. Consciência Situacional. 3. Gestão de Desastres. 4. Classificação Textual. 5. Identificação de Alagamentos. I. Ueyama, Jó, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Thiago Aparecido Gonçalves da Costa

Exploring a Multimodal Fusion approach to support Disaster Management: a case study with tweets and contextual data

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Jó Ueyama

USP – São Carlos
October 2021

Este trabalho é dedicado a Deus e a minha família.

AGRADECIMENTOS

Agradeço primeiramente a Deus que me deu saúde e forças para concluir os meus estudos na pós-graduação, além de demonstrar diariamente que é possível eu alcançar meus objetivos com muita fé, dedicação e esforço.

Agradeço aos meus pais Sueli e João pelo amor, incentivo e apoio incondicional para superar os desafios encontrados durante a minha trajetória de estudos no ICMC-USP. Além disso, gostaria de agradecer a minha irmã Karina e avó Francisca pelas orações e incentivo aos meus estudos na pós-graduação.

Agradeço ao professor Dr. Jó Ueyama pela oportunidade de ter sido seu orientado e estagiário, aliás, também sou grato pelos conselhos, atenção, paciência e orientação acadêmica.

Ao ICMC-USP (Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo) pela oportunidade de estudar em uma das melhores universidades da América Latina, além de fornecer toda a infraestrutura necessária para o desenvolvimento da minha pesquisa. Inclusive, gostaria de agradecer também a todos os funcionários administrativos e professores do ICMC-USP que me ajudaram no decorrer da minha pós-graduação, como: Kalinka Castelo Branco, Adenilso Simão, Solange Oliveira, Roseli Romero, Agma Traina, Francisco Rodrigues, Rudinei Goularte, Rodolfo Meneguette, Ricardo Marcacini, Maria da Graça Pimentel, Lycaena Couvre, Bruno Chaaban, entre outros.

Agradeço a Comissão de Pós-Graduação e de Relações Internacionais do ICMC-USP pela oportunidade de ter sido o Vice-Representante Discente durante o ano de 2019.

Ao Laboratório de Sistemas Web e Multimídia Interativos (Intermídia) pelo acolhimento, espaço e infraestrutura fornecida. Ademais, sou grato a todos os membros deste laboratório de pesquisa, pois acredito que a interação com eles foi crucial para uma boa estadia na cidade de São Carlos e a conclusão da minha pesquisa de mestrado, visto que vários alunos me auxiliaram durante o período de desenvolvimento deste trabalho.

Aos meus amigos da pós-graduação, Heitor Freitas, Erikson Aguiar, Alfredo Guilherme, Henrique Silva, Lucas Brito, Felipe Giuntini, Alef Vinícius, José Torres Neto, Kauê Moraes, Claudio Costa, Alyson de Jesus, Lucélia Santos, Geraldo Pereira, Rodrigo Kishi, Claudinei Brito, Renan Nespolo, Flávia Santos, Sandra Rodrigues, Marcelo Miky, Gustavo Escobedo, Matheus Takata, Sidgley Andrade, Wilk Oliveira, muito obrigado por toda ajuda, apoio e parcerias durante o meu período de estudos na pós-graduação. Além disso, um agradecimento especial ao professor Dr. Elias Adriano, pois ele me ajudou muito em um período difícil da pós-graduação, inclusive

sou grato por todos os conselhos, incentivo e apoio aos meus estudos e a minha carreira.

Agradeço aos professores, Dr. Elvis Fusco, Dr. Fábio Dacêncio, Dr. Rodolfo Chiaramonte e Dr. Leonardo Botega pelo incentivo e apoio ao meu ingresso no mestrado acadêmico após a conclusão da graduação em Ciências da Computação no UNIVEM (Centro Universitário Eurípedes de Marília).

Ao grupo de pesquisa AGORA (*A Geospatial Open collaboRative Architecture for building resilience against disasters and extreme events*) pela disponibilização da base de dados de mensagens publicadas no *Twitter* da cidade de São Paulo de novembro de 2016 até outubro de 2018, onde foi crucial para a execução desta pesquisa.

Agradeço a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) por confiar em minha pessoa e ter fornecido apoio financeiro para a execução desta pesquisa de mestrado (número de concessão PROEX-10839814/M).

“Educação não transforma o mundo.
Educação muda as pessoas.
Pessoas transformam o mundo.”
(Paulo Freire)

RESUMO

COSTA, T. A. G. **Explorando uma abordagem de Fusão Multimodal para auxílio a Gestão de Desastres: um estudo de caso com tweets e dados contextuais**. 2021. 193 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Atualmente, muito se tem discutido sobre a quantidade abundante de enchentes que assolam a cidade de São Paulo, inclusive esses desastres naturais preocupam excessivamente as autoridades governamentais paulistas, pois geram diversos prejuízos financeiros e sociais para a população afetada. Assim, existem iniciativas como a Gestão de Desastres que possuem o objetivo de prevenir e diminuir os impactos dos desastres naturais em nossa sociedade, visto que há fases com a tarefa de localizar e socorrer as vítimas das inundações. Dessa forma, as redes sociais (por exemplo, o *Twitter*) são essenciais para o auxílio da etapa de resposta da Gestão de Desastres, porque existe a disseminação de uma quantidade demasiada de mensagens relacionadas com alagamentos, no qual são capazes de serem úteis ao processo de localização de possíveis vítimas. No entanto, não é uma tarefa simples a obtenção da Consciência Situacional de desastres naturais a partir de *tweets*, visto que essas mensagens são frequentemente escritas de maneira coloquial e os algoritmos de Aprendizado de Máquina não são capazes de compreender o contexto das mensagens do *Twitter*. Dessa forma, com o objetivo de melhorar o processo de classificação textual e consequentemente captar Consciência Situacional de alagamentos de maneira mais precisa, então esta pesquisa investiga a Fusão Multimodal de informações textuais com contextuais. Esta pesquisa tem o objetivo de desenvolver uma abordagem de Fusão Multimodal capaz de auxiliar a etapa de resposta da Gestão de Desastres a partir de *tweets*, dados climáticos e incidências históricas de enchentes, além de implementar um *software* capaz de detectar possíveis vítimas de enchentes. Em vista disso, foram desenvolvidos mecanismos computacionais capazes de realizar o Processamento de Linguagem Natural das mensagens do *Twitter*, descobrir as regiões propícias ao acontecimento de alagamentos da capital paulista e combinar os dados heterogêneos por intermédio de estratégias baseadas em Aprendizado de Máquina. Os resultados revelam que o modelo de Fusão Multimodal do tipo híbrido com foco na decisão proporcionada pelos dados meteorológicos pode identificar as possíveis vítimas de alagamentos com 84,70% de precisão, aliás combinar dados textuais com multimodais proporciona um acréscimo de 18,53% na precisão da obtenção de Consciência Situacional de inundações, portanto para auxílio a fase de resposta da Gestão de Desastres, abordagens multimodais são mais eficazes dos que as unimodais. Ademais, algoritmos de agrupamento hierárquico demonstraram ser capazes de descobrir regiões propícias ao acontecimento de enchentes da cidade de São Paulo mais bem definidas do que os mecanismos de agrupamento baseados em densidade. Além disso, estratégias de definição da distância máxima de formação de áreas de alagamentos embasadas em abordagens empíricas se mostraram mais promissoras do

que as baseadas em estratégias geo estatísticas. Por último, esta abordagem de Fusão Multimodal pode ser adaptada para diferentes idiomas, regiões e desastres naturais, além de que o *software* desenvolvido pode auxiliar as autoridades governamentais a localizar possíveis vítimas de inundações da capital paulista em tempo real.

Palavras-chave: Fusão Multimodal, Consciência Situacional, Gestão de Desastres, Classificação Textual, Identificação de Alagamentos.

ABSTRACT

COSTA, T. A. G. **Exploring a Multimodal Fusion approach to support Disaster Management: a case study with tweets and contextual data**. 2021. 193 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Currently, much has been discussed about the abundant amount of floods that plague the city of São Paulo, including these natural disasters that worry the São Paulo government authorities excessively because they generate several financial and social losses for the affected population. Thus, there are initiatives such as Disaster Management that aim to corroborate the prevention and reduction of the impacts of natural disasters in our society since there are phases to locate and assist the victims of flooding. Therefore, social networks (for example, Twitter) are essential to help in the response stage of Disaster Management because there is the dissemination of too many messages related to floods, which are able to be useful in the process of locating possible victims. However, obtaining Situational Awareness of natural disasters from tweets is not a simple task since these messages are often written colloquially, and Machine Learning algorithms cannot understand the context of Twitter messages. Thus, to improve the textual classification process and consequently capture Situational Awareness of flooding more accurately, this research investigates the Multimodal Fusion of textual with contextual information. This research aims to develop a Multimodal Fusion approach capable of assisting the response stage of Disaster Management from tweets, climate data, and historical flood incidences and implementing software capable of detecting potential flood victims. Therefore, computational mechanisms capable of performing Natural Language Processing of the Twitter messages, discovering the regions prone to flooding in the capital city of São Paulo, and combining the heterogeneous data through strategies based on Machine Learning were developed. The results reveal that the hybrid type Multimodal Fusion model focusing on the decision provided by meteorological data can identify the possible victims of flooding with 84.70% accuracy. Moreover, combining textual data with multimodal data provides an 18.53% increase in the accuracy of obtaining Situational Awareness of flooding, so to aid the response phase of Disaster Management, multimodal approaches are more effective than unimodal ones. Furthermore, hierarchical clustering algorithms proved to discover better-defined flood-prone regions in the city of São Paulo than density-based clustering mechanisms; moreover, strategies for defining the maximum distance for flood area formation based on empirical approaches showed to be more promising than those based on geo-statistical strategies. Finally, this Multimodal Fusion approach can be adapted to different languages, regions, and natural disasters. In addition, the developed software can help government authorities locate possible victims of flooding in São Paulo in real-time.

Keywords: Multimodal Fusion, Situational Awareness, Disaster Management, Text Classification, Identification of Floods.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AD	Árvore de Decisão
AGORA	*A Geospatial Open collaboRative Architecture for building resilience against disasters and extreme events*
AM	Aprendizado de Máquina
API	*Application Programming Interface*
BOW	*Bag Of Words*
CBOW	*Continuous Bag of Words*
CGE-SP	Centro de Emergências Climáticas da Prefeitura de São Paulo
CNN	*Convolutional Neural Network*
CS	Consciência Situacional
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
DC	Defesa Civil
DL	*Deep Learning*
DST	*Dempster-Shafer Theory*
DT	*Decision Tree*
ER	Entidade Relacionamento
EUA	Estados Unidos da América
GBT	*Gradient Boosted Tree*
GD	Gestão de Desastres
IA	Inteligência Artificial
ICMC	Instituto de Ciências Matemáticas e de Computação
INMET	Instituto Nacional de Meteorologia
JMA	*Japan Meteorological Agency*
JSON	*JavaScript Object Notation*
KNN	*K-Nearest Neighbors*
LR	*Logistic Regression*
MD	Mineração de Dados
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
MT	Mineração de Texto
NB	*Naive Bayes*

NILC	Núcleo Interinstitucional de Linguística Computacional
NLTK	*Natural Language Toolkit*
NN	*Neural Networks*
NY	Nova York
ONGs	Organizações não Governamentais
OPTICS	*Ordering Points To Identify the Clustering Structure*
PAE	Programa de Aperfeiçoamento de Ensino
PLN	Processamento de Linguagem Natural
RF	*Random Forest*
RL	Regressão Logística
RNA	Redes Neurais Artificiais
RSSF	Redes de Sensores Sem Fio
SAW	*Situational Awareness*
SIFT	*Scale-invariant feature transform*
ST-DBSCAN	*Spatio-Temporal Density-Based Spatial Clustering of Applications with Noise*
SVM	*Support Vector Machine*
TAGGS	*Toponym-based Algorithm for Grouped Geoparsing of Social media*
TF-IDF	*Term Frequency-Inverse Document Frequency*
TL	*Transfer Learning*
TM	*Text Mining*
URLs	*Uniform Resource Locators*
USP	Universidade de São Paulo
VGI	*Volunteered Geographic Information*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Neste capítulo é apresentado a introdução desta dissertação, assim na Seção 1.1 observa-se a contextualização e a motivação desta pesquisa de mestrado, em seguida na Seção 1.2 são expostos os objetivos deste trabalho, já na Seção 1.3 apresenta-se a organização do texto desta dissertação.

1.1 Contextualização e Motivação

Hoje em dia, há uma quantidade exacerbada de ocorrências de enchentes, alagamentos e inundações em território brasileiro, inclusive esses fenômenos naturais são responsáveis pelos diversos prejuízos ocasionados na qualidade de vida da população afetada. Por exemplo, os alagamentos que acontecem na cidade de São Paulo que adicionados aos problemas de infraestrutura pré-existentes do município, geram os seguintes transtornos para a sociedade: prejuízos econômicos, desperdícios materiais, uma parte da população desabrigada e mortes (HADDAD; TEIXEIRA, 2015).

A princípio, as enchentes são fenômenos naturais caracterizados pelo acréscimo do nível da água de um canal até seu limite (CASTRO *et al.*, 1996). Ademais, as inundações ocorrem a partir do transbordamento da água de um determinado canal devido às chuvas intensas em regiões de relevo acidentado (CASTRO *et al.*, 1996). Já o surgimento dos alagamentos, relaciona-se com o acúmulo de água advinda de chuvas intensas em perímetros urbanos com sistemas de drenagem ineficazes (AMARAL; GUTJAHR, 2011).

Desse modo, as consequências que esses fenômenos naturais oferecem para a população afetada são: danos humanos e materiais; pausa de atividades econômicas das regiões alagadas; doenças de propagação hídrica (por exemplo, cólera, leptospirose, giardíase, entre outras); poluição da água por materiais tóxicos (TUCCI; BERTONI *et al.*, 2003). Inclusive, as enchentes, inundações e alagamentos são definidos como “desastres naturais”, visto que esses eventos

exorbitantes acontecem em locais que são habitados por seres humanos (KOBIYAMA *et al.*, 2006).

Além disso, segundo Mizutori e Guha-Sapir (2017) no período de 1997 até 2017 os alagamentos, inundações e enchentes proporcionaram prejuízos econômicos de 656 bilhões de dólares em todo globo terrestre, além de que durante esses 20 anos esses desastres naturais foram responsáveis pela morte de 142.088 pessoas em todo o mundo. Inclusive, de acordo com Haddad e Teixeira (2015) as consequências financeiras anuais que os alagamentos proporcionam para a capital do estado de São Paulo são de R$218,19 milhões de reais, assim as inundações contribuem para a redução do crescimento do município e a diminuição da qualidade de vida dos habitantes da capital paulista, além de prejudicar a competitividade econômica das empresas locais e internacionais que atuam naquela região.

Dado o exposto, existem atividades como a Gestão de Desastres (GD) que possuem o intuito de evitar ou minimizar o impacto dos desastres naturais na sociedade, além de oferecer auxílio e proporcionar a resiliência, adaptação e transferência dos locais de residência dos moradores prejudicados pelos fenômenos naturais (NORRIS *et al.*, 2008; BAHARIN; SHIBGHATULLAH; OTHMAN, 2009; MENDIONDO, 2010; POSER; DRANSCH, 2010). Desse modo, para a GD as fontes de dados heterogêneas (por exemplo, imagens, redes sociais, sensores, radares, entre outras) são extremamente importantes para a minimização dos prejuízos ocasionados pelos desastres naturais nas populações afetadas, visto que as autoridades são capazes de emitir notificações para a Defesa Civil (DC), agências de emergência e organizações humanitárias (HORITA *et al.*, 2015).

Na pesquisa de Hughes *et al.* (2011), observa-se uma abordagem que possui o objetivo de auxiliar a GD e que usufrui exclusivamente de dados advindos de sensores, logo para o desenvolvimento do trabalho os autores aplicaram a tecnologia de Redes de Sensores Sem Fio (RSSF) atrelada a dispositivos embarcados capazes de monitorar o nível da água dos rios e sua velocidade de acréscimo em períodos de chuvas, com o intuito de notificar as autoridades sobre a possibilidade da ocorrência de alagamentos em algumas regiões da cidade de São Carlos (HUGHES *et al.*, 2011). No entanto, essa estratégia de apoio a GD é suscetível a diversos erros, pois os dispositivos embarcados são passíveis a falhas de conexão com os agentes computacionais externos, os equipamentos podem ser furtados e os dispositivos embarcados podem apresentar problemas no *hardware*, no qual são ocasionados pela exposição ao meio ambiente. Portanto, necessita-se o desenvolvimento de abordagens computacionais que usufruam de outras mídias para o auxílio efetivo a GD.

Atualmente, as redes sociais são responsáveis pela produção de uma quantidade acentuada de dados relevantes sobre fenômenos naturais (WIN; AUNG, 2017), além de serem uma fonte de informação que pode ser explorada para a detecção, monitoramento e predição de diversos eventos (STEIGER; ALBUQUERQUE; ZIPF, 2015). Inclusive, nos últimos anos os pesquisadores ampliaram o interesse pela colaboração em pesquisas que possuem o objetivo de analisar as

mensagens advindas de redes sociais com o intuito de auxiliar a GD (ALBUQUERQUE *et al.*, 2015).

Dessa forma, nota-se na literatura uma tendência na publicação de artigos em conferências e periódicos que utilizem a análise das mensagens do *Twitter*[1] para a identificação de fenômenos naturais, como terremotos (por exemplo, Sakaki, Okazaki e Matsuo (2010), Earle, Bowden e Guy (2012), Crooks *et al.* (2013), Avvenuti *et al.* (2014)), incêndios florestais (por exemplo, Longueville, Smith e Luraschi (2009), Spinsanti e Ostermann (2013)), furacões (por exemplo, Kryvasheyeu *et al.* (2016)) e inundações (por exemplo, Brouwer *et al.* (2017)). Dessa maneira, o *Twitter* é a rede social selecionada por essas pesquisas em razão da facilidade de captação de informações públicas, há uma quantidade exorbitante de mensagens disseminadas diariamente, inclusive de acordo com Kwak *et al.* (2010), inúmeras notícias relacionadas com desastres naturais são divulgadas primeiramente no *Twitter* do que nos veículos de comunicação tradicionais (por exemplo, televisão, rádio, entre outros).

O *Twitter* proporciona aos usuários o compartilhamento de inúmeras mensagens curtas com o tamanho de até 280 caracteres. Além disso, segundo Bruijn *et al.* (2020) diariamente há a publicação de mais de 500 milhões de mensagens, sendo que 20.000 *tweets* possuem palavras relacionadas com desastres naturais (por exemplo, "*flood*", onde a tradução para o português corresponde a "alagamento"). Dessa forma, a extração, monitoramento e análise das mensagens publicadas no *Twitter* proporciona indícios das possíveis localizações de desastres naturais, descoberta de possíveis vítimas e outros dados relevantes para as organizações de apoio a emergências (BRUIJN *et al.*, 2018).

Entretanto, os usuários do *Twitter* frequentemente compartilham mensagens escritas de forma coloquial, utilizam abreviações, usufruem de metáforas para se referenciarem a algumas palavras relacionadas a desastres naturais (por exemplo, "chuva", no qual os usuários se expressam como "chuva de bençãos" e "inundação" como "a mente está inundada de ideias") e compartilham *Volunteered Geographic Information* (VGI), em português Informações Geográficas Voluntárias, errôneas intencionalmente ou involuntariamente (YIN *et al.*, 2012; FENG; SESTER, 2018). Portanto, é necessário um processo de filtragem textual efetivo para que seja possível aprimorar o desempenho dos algoritmos de Aprendizado de Máquina (AM) utilizados no Processamento de Linguagem Natural (PLN), onde possui a função de captar a *Situational Awareness* (SAW), em português Consciência Situacional (CS), de desastres naturais a partir de *tweets*. Não menos importante, SAW é caracterizado pela capacidade de um indivíduo compreender uma determinada situação (ENDSLEY, 1988).

De acordo com pesquisas realizadas na literatura, observou-se que há uma quantidade expressiva de trabalhos focados na captação da SAW de fenômenos naturais a partir de *tweets*, aliás esses trabalhos propõem abordagens computacionais que utilizam diversos algoritmos de AM para a classificação das informações, como *Random Forest* (RF), *Decision Tree* (DT), *Support*

[1] url: <www.twitter.com>

Vector Machine (SVM), *Naive Bayes* (NB), *Neural Networks* (NN), entre outros (BRUIJN *et al.*, 2020). Porém, esses algoritmos ao serem utilizados para o PLN apresentam discrepâncias nos resultados ao serem comparados com a interpretação de texto realizada pelos seres humanos, pois as pessoas são capazes de usufruir do contexto para a compreensão das mensagens (BRUIJN *et al.*, 2020). Além disso, utilizar somente dados textuais nos mecanismos de classificação proporciona a recuperação de diversos falsos positivos, visto que é um espaço amostral com ambiguidades inerentes (FENG; SESTER, 2018). Assim, para aprimorar o processo de classificação de informações textuais nota-se na literatura a utilização da estratégia de Fusão Multimodal, no qual é caracterizada pela capacidade de combinar diversas informações midiáticas para a realização de atividades específicas (XIE; GUAN, 2013). Inclusive, observa-se na literatura poucos trabalhos que utilizam diversas fontes de dados para aperfeiçoar o processo de classificação texto com o intuito de auxiliar a GD (ALBUQUERQUE *et al.*, 2015).

Por último, a hipótese desta pesquisa é que a **Fusão Multimodal de informações textuais, dados meteorológicos e localizações históricas de alagamentos, produza um modelo de obtenção de SAW de desastres naturais que é mais preciso para o auxílio da etapa de resposta da GD do que abordagens unimodais**. Sendo que, as informações meteorológicas contribuem para o modelo de Fusão Multimodal, pois existem diversos *tweets* com VGI errôneas e várias mensagens contendo metáforas, assim para a solucionar esses empecilhos é necessário a combinação de dados textuais com informações contextuais. Já as localizações históricas de alagamentos, possuem a tendência de corroborar com o modelo de Fusão Multimodal, pois de acordo com Tobler (1970) a probabilidade de uma coordenada geográfica ser correlata a outra é inversamente proporcional ao seu distanciamento, ou seja, usuários que publicaram mensagens relacionadas a alagamentos e são possíveis vítimas de desastres naturais tendem estar próximos dos fenômenos do que distantes. Aliás, nota-se na literatura que há a ausência de pesquisas que usufruam de informações geográficas contextuais para melhorar a classificação textual em abordagens de Fusão Multimodal (BRUIJN *et al.*, 2020).

1.2 Objetivos

Esta dissertação de mestrado possui o objetivo de desenvolver um modelo de Fusão Multimodal para obter a SAW de inundações e auxiliar a etapa de resposta da GD, inclusive este modelo usufrui de informações textuais advindas de *tweets*, dados meteorológicos e ocorrências históricas de alagamentos da cidade de São Paulo. Além disso, existe a descrição do desenvolvimento de um *software* de identificação de possíveis vítimas de alagamentos da cidade de São Paulo, onde utiliza do modelo de Fusão Multimodal elaborado nesta pesquisa para a obtenção de SAW de enchentes. Não menos importante, o desenvolvimento deste *software* é motivado pela possibilidade de proporcionar aos moradores da cidade de São Paulo a oportunidade de minimizar os impactos das inundações em suas vidas.

Desse modo, com o intuito de alcançar o objetivo principal proposto, assim há uma série de objetivos específicos definidos:

- Definir qual a melhor estratégia para a identificação das áreas de alagamentos da cidade de São Paulo a partir dos dados históricos de ocorrências de inundações;
- Determinar qual a melhor abordagem para a definição da distância de formação dos agrupamentos (empírica ou estatística);
- Explorar as diversas categorias de Fusão Multimodal, como: Prévia, Tardia, Híbrida com decisão nos dados textuais e Híbrida com decisão nas informações meteorológicas;
- Reconhecer o impacto no desempenho dos modelos de obtenção de SAW de desastres naturais quanto a inclusão de dados meteorológicos e geográficos.
- Desenvolver o *software* que identifica possíveis vítimas de alagamentos em tempo real a partir de um modelo de Fusão Multimodal que utiliza informações contextuais (ou seja, dados textuais, informações meteorológicas e ocorrências históricas de enchentes da cidade de São Paulo).

1.3 Organização do Texto

Esta dissertação foi organizada da seguinte maneira. O Capítulo 2 expõe o referencial teórico desta pesquisa, além de exibir alguns conceitos empregados no decorrer do trabalho. Já o Capítulo 3, apresenta e sintetiza os trabalhos correlatos desta pesquisa, aliás realiza-se uma comparação entre as pesquisas relacionadas e a abordagem elaborada nesta dissertação. O Capítulo 4 exibe a metodologia empregada para o desenvolvimento do modelo de Fusão Multimodal, no qual pode obter SAW de alagamentos e auxiliar a etapa de resposta da GD, além disso, há a exposição do planejamento dos experimentos para a criação e validação do modelo. Já o Capítulo 5, apresenta a arquitetura informacional da plataforma de detecção de possíveis vítimas de alagamentos, além de detalhar as funcionalidades das camadas do *software*. Já o Capítulo 6, apresenta as métricas de avaliação utilizadas nos experimentos, além de discutir os resultados dos experimentos empregados para criar e avaliar a abordagem proposta nesta dissertação, sendo que neste capítulo observam-se experimentos que vão desde a descoberta das regiões propícias ao acontecimento de inundações na cidade de São Paulo até a comparação entre os diversos modelos de Fusão Multimodal. Por último, o Capítulo 7 reporta as conclusões desta pesquisa e exibe quais são as principais contribuições científicas, além disso, há a apresentação de quais são os trabalhos futuros e as atividades acadêmicas e complementares realizadas pelo aluno no decorrer da pós-graduação.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos que serão utilizados neste trabalho. Primeiramente, é discutido sobre Gerenciamento de Desastres Naturais (Seção 2.1), Consciência Situacional (Seção 2.2) e Fusão Multimodal (Seção 2.3). Posteriormente, é explorado a metodologia inerente aos processos de Mineração de Texto (Seção 2.4) e *Word Embeddings* (Seção 2.5). Por último, destaca-se as particularidades dos algoritmos de AM mais utilizados na literatura para trabalhos envolvendo a obtenção de SAW de fenômenos a partir de mensagens publicadas em redes sociais e a definição de zonas propícias a alagamentos (Seção 2.6), além de apresentar as características da área de Geoestatística e da técnica capaz de gerar um modelo de dependência espacial das ocorrências históricas de alagamentos (Seção 2.7).

2.1 Gerenciamento de Desastres Naturais

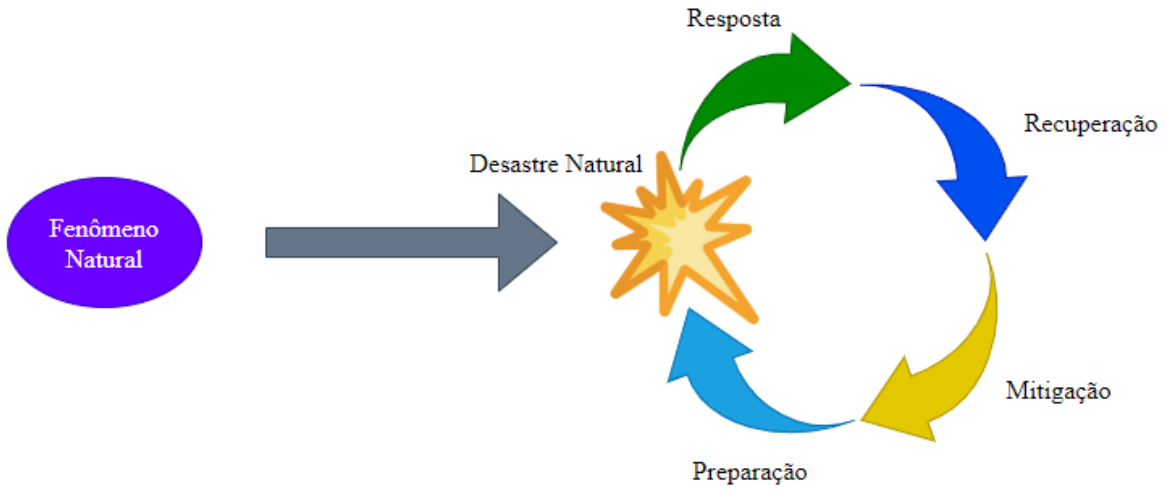
Nas últimas décadas, os desastres naturais aumentaram a intensidade e a frequência de maneira exorbitante, causando diversos prejuízos para a sociedade, como: danos materiais, socioeconômicos e mortes (CUTTER; EMRICH, 2005; KOBIYAMA *et al.*, 2006; KLOMP, 2016). Inclusive, desastres naturais são fenômenos naturais extremos que afetam de forma econômica, social e estrutural uma sociedade, onde os impactos são tão grandes que excedem a capacidade da comunidade gerir seus próprios recursos (LONGUEVILLE *et al.*, 2010).

Além disso, segundo UNISDR (2009) um desastre pode ser mensurado a partir da análise de quão exposta, vulnerável e capaz de lidar com fenômenos naturais uma comunidade está. Ou seja, a **exposição** é relacionada com a quantidade de fenômenos naturais propícios a acontecer na região, já a **vulnerabilidade** é contabilizada a partir de uma análise de quantas áreas de risco existentes e quantas pessoas residem nelas, por último, a **capacidade** é examinada a partir da quantidade de mecanismos de gerenciamento de desastres naturais existentes na região estudada (por exemplo, alarmes de emergência, sistemas de drenagem, entre outros) (UNISDR, 2009).

De acordo com o IBGE (2010), somente na cidade de São Paulo existem 674.329 pessoas vivendo em áreas de risco de ocorrência de desastres naturais, como: enchentes, alagamentos, inundações, deslizamentos de terra, entre outros. Portanto, para a redução dos impactos dos desastres, necessita-se de uma Gestão de Desastres, pois essa medida garante a resiliência, adaptação e mudança das comunidades afetadas (NORRIS *et al.*, 2008; BAHARIN; SHIBGHATULLAH; OTHMAN, 2009; MENDIONDO, 2010; POSER; DRANSCH, 2010).

O objetivo da GD é reduzir ou evitar prejuízos à sociedade e proporcionar suporte as vítimas de fenômenos naturais extremos (POSER; DRANSCH, 2010). A GD é um processo contínuo e composto de diversas etapas executadas antes, durante e depois da ocorrência de fenômenos naturais, além disso, é divida em quatro partes: **mitigação**, **preparação**, **resposta** e **recuperação** (HORITA, 2017). Na GD, todas as fases possuem diferentes necessidades informacionais para melhorar a realização das atividades desejadas, portanto é preciso que elas sejam precisas e constantemente atualizadas (DEGROSSI, 2015). A seguir, na Figura 1 encontra-se os ciclos da GD.

Figura 1 – Fases da Gestão de Desastres



Fonte: Adaptada de Poser e Dransch (2010).

Na lista a seguir serão detalhadas as etapas presentes na Figura 1:

- **Mitigação:** Esta etapa caracteriza-se pela reunião de atividades com o intuito de reduzir a vulnerabilidade de uma comunidade e minimizar os impactos de futuros desastres naturais (VIVACQUA; BORGES, 2012). Segundo Vivacqua e Borges (2012), a eficiência das atividades presentes nesta etapa dependerão da disponibilidade de informações referentes a riscos, perigos e medidas que serão tomadas;
- **Preparação:** De acordo com Poser e Dransch (2010), Vivacqua e Borges (2012), esta etapa caracteriza-se pelo planejamento da resposta aos desastres naturais, logo, inclui o

treinamento de equipe de emergência, avaliação contínua do planejamento, instalação de equipamentos de monitoramento, alerta e predição de desastres;

- **Resposta:** Esta etapa ocorre após o desastre acontecer, dessa forma ela tem algumas particularidades, como: imprevisibilidade, os acontecimentos ocorrem rapidamente, os recursos são indisponíveis, ações e decisões precisam ser rápidas e eficientes, além de que há uma quantidade de pessoas afetadas (VIVACQUA; BORGES, 2012). Aliás, segundo Poser e Dransch (2010) as medidas de resposta visam manter ou reestabelecer a segurança pública, realizar operações de busca e salvamento e oferecer ajuda humanitária para a população afetada;
- **Recuperação:** Esta fase ocorre após o controle do desastre, logo é um conjunto de atividades para reparar, reconstruir e recuperar o que foi perdido pela população no desastre natural (POSER; DRANSCH, 2010).

Para a GD, o monitoramento de fontes de dados heterogêneas (por exemplo, sensores, radares, redes sociais, entre outras) é crucial para a redução dos impactos dos desastres naturais, pois as autoridades governamentais podem identificar os fenômenos a partir dos sistemas disponíveis e emitir antecipadamente notificações para a DC, agências de emergência e organizações humanitárias para que essas instituições auxiliem os residentes das áreas de risco (HORITA *et al.*, 2015).

2.2 Consciência Situacional

A SAW, na Língua Portuguesa é chamada de Consciência Situacional, é definida por Endsley (1988) como: “a identificação dos objetos num determinado tempo e espaço, percepção de seus significados e a idealização de seus estados em um futuro iminente”. Ou seja, SAW é um processo cognitivo dos seres humanos, onde o indivíduo pode ver, analisar e compreender o ambiente ou a circunstância em que ele se encontra (ENDSLEY, 1988). Desse modo, de acordo com Roy e Wark (2007), SAW é um processo que auxilia na tomada de decisão humana.

Além disso, segundo Botega (2016) SAW está presente em todas as circunstâncias da vida dos seres humanos, essas que vão desde a compreensão de situações operacionais até acontecimentos corriqueiros. Por exemplo, um jogador necessita avaliar as condições em um campo de futebol em uma partida, já um motorista precisa avaliar todos os elementos presentes no ambiente dinâmico que é a estrada, caso deseje uma ultrapassagem, além de que um mergulhador precisa entender quais são as circunstâncias de mergulho para que ele se mantenha seguro no mar (BOTEGA, 2016).

A aeronáutica é um domínio que exige altíssimos níveis de SAW, logo foi uma das primeiras áreas que adotaram os conceitos de SAW para auxiliar nas tarefas dos tomadores de decisão. Outrora, este processo não fica restrito ao ambiente militar, sendo aplicados em outros,

como: gestão de tráfego aéreo e urbano, laudo médico, análise de dados da bolsa de valores, gerenciamento de desastres naturais, entre outros. Portanto, aplica-se SAW em situações que necessitam de tomadas de decisão rápidas e que são relacionadas com a vida dos seres humanos e bens materiais (ENDSLEY, 2001; KOKAR; ENDSLEY, 2012).

A SAW é o resultado do processo cognitivo humano, ou seja, a representação do conhecimento fundamental do estado de um elemento e de suas mudanças num determinado espaço e tempo (ENDSLEY, 2016). Dessa forma, segundo Botega (2016), com o auxílio da SAW, os profissionais responsáveis pelas tomadas de decisões em ambientes críticos conseguem decidir quais ações devem ser tomadas nas circunstâncias em que se encontram, com o intuito de obterem êxito em suas tarefas. Entretanto, uma ótima SAW não garante uma ótima tomada de decisão, por isso que os sistemas computacionais que proporcionam SAW devem ser constantemente aperfeiçoados para que seja aumentada a probabilidade da execução das ações apropriadas em determinadas circunstâncias pelos tomadores de decisão (BOTEGA, 2016).

2.2.1 Modelo de Consciência Situacional de Endsley

Atualmente, dentro os modelos de SAW existentes na literatura, um dos modelos mais utilizados é o de Endsley (1988), no qual é empregado por pesquisadores para aderir SAW em sistemas computacionais que possuem o objetivo de apoiar a tomada de decisão em ambientes críticos. De acordo com Endsley (1988), SAW é um modelo composto por um grupo de componentes relacionados em três níveis, no qual são influenciados por tarefas e sistemas de maneira interna e externa.

A seguir serão discutidos os três principais níveis de SAW observados na Figura 2:

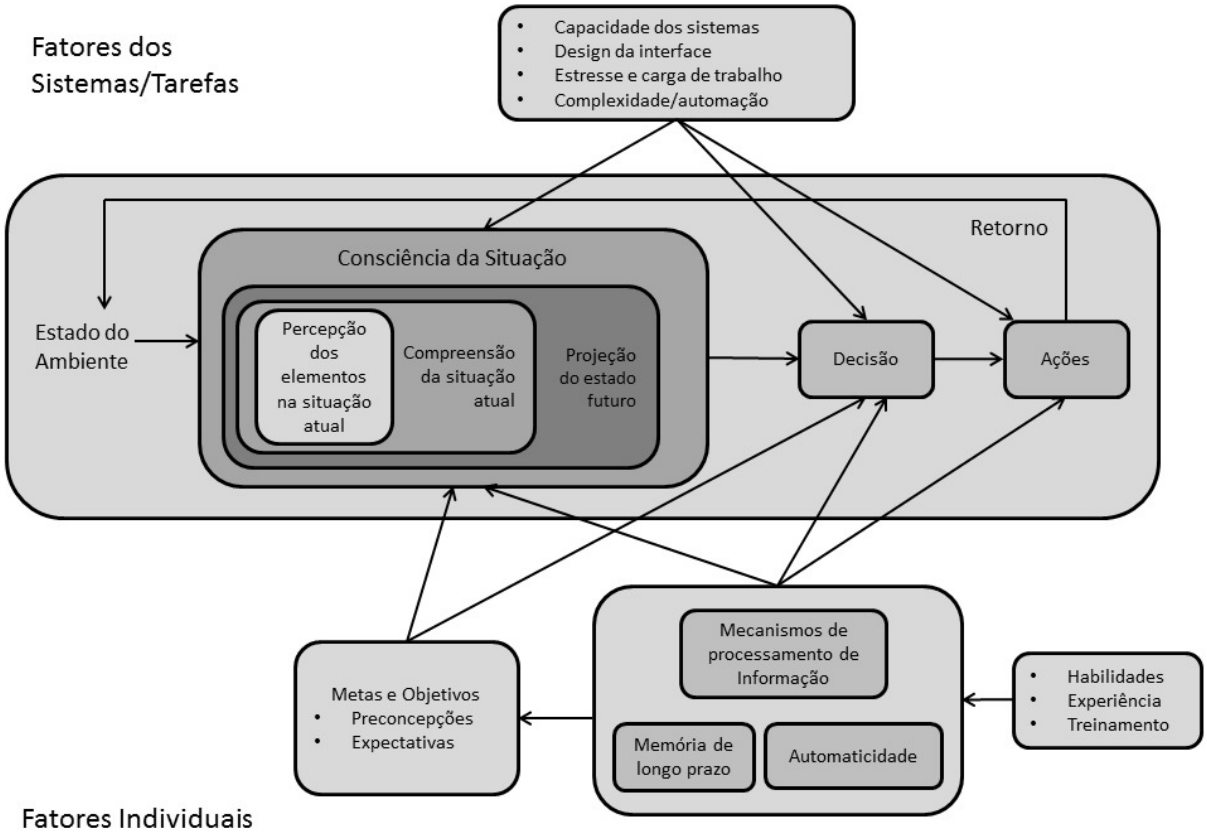
1. **Percepção dos elementos no ambiente**: O primeiro nível de SAW, possui o objetivo de notar os elementos presentes num determinado ambiente, suas características e estados, de maneira visual, textual ou sonora (BOTEGA, 2016);

2. **Compreensão da situação atual**: O segundo nível de SAW, é responsável pela percepção dos elementos no ambiente e quais são as relações que eles possuem entre eles, considerando a importância que esses elementos possuem em cumprir objetivos e metas de uma situação (BOTEGA, 2016). Inclusive, não é possível projetar os elementos em um futuro imediato sem ter compreendido a situação atual deles (ENDSLEY, 1988);

3. **Projeção dos estados futuros**: De acordo com Botega (2016), o terceiro nível de SAW, destaca-se pela percepção do que os elementos significam no estado atual e assim predizer quais são suas próximas ações perante ao objetivo principal.

Dado o exposto, de acordo com Endsley (1988), as entradas para o núcleo do modelo de SAW são os sinais ou estados obtidos do ambiente analisado, desse modo o modelo de SAW

Figura 2 – Modelo de SAW de Endsley



Fonte: Botega (2016), Endsley (1988).

proporciona aos tomadores de decisão a percepção, projeção e compreensão dos elementos daquele ambiente. Ademais, segundo Botega (2016) o módulo de decisão é fora do núcleo de SAW propositalmente, pois SAW é um modelo mental do tomador de decisão sobre a condição que se encontram os elementos no ambiente, logo quem decide qual ação deve ser tomada dada as circunstâncias do espaço é o operador.

O núcleo de SAW é influenciado diretamente por características externas dos elementos, das atividades, dos sistemas computacionais e do ambiente físico em que está sendo analisado (BOTEGA, 2016). Inclusive, segundo Botega (2016) os sistemas computacionais influenciam o núcleo de SAW por intermédio das suas áreas de interação e capacidades, do quão complexo e automatizado se encontra suas tarefas, e da quantidade de estresse e carga de trabalho que proporciona nos operadores. Além disso, os atributos do ambiente que são capazes de induzir diretamente a SAW são: ruídos, iluminação, chuva, ventania, temperatura, entre outras.

Os fatores externos advindos de sistemas ou do ambiente podem facilitar, ou dificultar a obtenção de SAW do tomador de decisão, visto que esse indivíduo influenciará diretamente o núcleo do modelo de SAW através de suas experiências e habilidades de execução das tarefas necessárias (BOTEGA, 2016). Inclusive, o papel do módulo de objetivos e metas é encaminhar a

atenção do operador aos sinais presentes no ambiente analisado (BOTEGA, 2016).

Por último, de acordo com Botega (2016) os mecanismos com a função de processar a informação possuem o objetivo de prover estratégias para competir pela atenção do tomador de decisão diante das diversas fontes de informação presentes no ambiente. Já o módulo de automaticidade, induz a SAW a partir da aplicação automática e repetida das atividades de rotina dos operadores (BOTEGA, 2016).

2.3 Fusão Multimodal

A Fusão Multimodal é uma estratégia que combina diversas mídias (por exemplo, áudios, vídeos, textos, sensores, entre outras) com o objetivo de realizar atividades específicas, como a identificação de fenômenos naturais, localização de seres humanos, identificação de contextos semânticos em mensagens publicadas em redes sociais e detecção de locutores através de informações audiovisuais (XIE; GUAN, 2013). Inclusive, o conceito de multimodalidade intrínseco nas estratégias de Fusão Multimodal está relacionado com a habilidade de um sistema computacional interagir com os seus usuários, a partir de diversos canais de comunicação, além de usufruir de técnicas de extração e transmissão de significado automaticamente (NIGAY; COUTAZ, 1993).

A Fusão de dados midiáticos, sinônimo de Fusão Multimodal, de acordo com Atrey *et al.* (2010) possui alguns obstáculos que devem ser resolvidos ou minimizados, como:

- As informações midiáticas possuem distintos graus de confiança ao serem utilizadas para a execução de algumas atividades;
- As mídias podem ser correlatas ou independentes, onde a correlação é benéfica para a validação dos processos de decisão, já a independência é importante para a complementação dos demais dados;
- Os formatos e as proporções das informações midiáticas são distintos;
- As durações de processamento das diversas categorias de mídia são diferentes.

Dado o exposto, segundo Atrey *et al.* (2010) para a solução ou minimização dos obstáculos apresentados, as estratégias de Fusão Multimodal devem responder alguns questionamentos relevantes, como:

1. Qual o nível da fusão executada?
2. Como será realizado o processo de fusão?
3. Quando será executada a fusão?

4. Quais as categorias de mídias que serão combinadas?

O primeiro questionamento é referente ao nível da Fusão Multimodal, assim as possibilidades de combinação de dados midiáticos são: ***early fusion*** (na Língua Portuguesa é conhecida como Fusão Prévia), ***late fusion*** (na Língua Portuguesa é conhecida como Fusão Tardia) e ***hybrid fusion*** (na Língua Portuguesa é conhecida como Fusão Híbrida) (LOPES, 2015). Inclusive, essas categorias de Fusão Multimodal serão descritas em Subseção 2.3.1, Subseção 2.3.2, Subseção 2.3.3, respectivamente.

Já a segunda pergunta, de acordo com Lopes (2015) relaciona-se a escolha de uma metodologia adequada para a combinação das diferentes mídias, logo os métodos são classificados em: união das informações midiáticas através de regras (por exemplo, combinação linear); tratamento da problemática de fusão de características heterogêneas como um desafio de classificação de informações por intermédio de algoritmos de AM (por exemplo, *Random Forest*, *Decision Tree*, entre outros); por último, fusão dos dados multimodais com o intuito da realização de previsões dos resultados. Inclusive, neste trabalho foram empregadas as metodologias da primeira e segunda categoria.

O terceiro questionamento é referente a escolha do período adequado para a execução do processo de Fusão Multimodal, visto que o primeiro obstáculo que deve ser solucionado é referente as diferentes proporções de amostragens, ou seja, desbalanceamento da base de dados, já o segundo o desafio é referente a sincronização das diversas características de diferentes mídias (LOPES, 2015). Dessa forma, para a solução do primeiro obstáculo deste questionamento, necessita-se balancear as taxas de informações midiáticas (por exemplo, em problemas solucionáveis com classificação binária utiliza-se 50% dos dados negativos e 50% dos dados positivos para não enviesar o classificador) (FACELI *et al.*, 2011). Outrora, a solução do segundo obstáculo é referente a dificuldade de sincronização de dados multimodais, logo uma das estratégias comumente empregadas na literatura é a fusão de informações temporalmente e geograficamente (por exemplo, combinação de dados de sensores e *tweets* da mesma cidade ou bacia hidrográfica e de um período de postagem semelhante) (ANDRADE *et al.*, 2017; BRUIJN *et al.*, 2020).

Por último, a quarta pergunta é referente a escolha das características mais adequadas das informações midiáticas para o processo de Fusão Multimodal, pois características diferentes podem exibir dados divergentes (LOPES, 2015). Dessa forma, existem diversas maneiras de seleção dessas características, sendo que uma das alternativas mais comumente empregada em trabalhos encontrados na literatura é a busca pelos atributos mais utilizados em pesquisas correlatas, por exemplo, em trabalhos que combinam informações textuais e meteorológicas os dados dos sensores de precipitação e a quantia de chuva acumulada em um determinado intervalo de tempo são frequentemente utilizados (YERVA; JEUNG; ABERER, 2012; BRUIJN *et al.*, 2020).

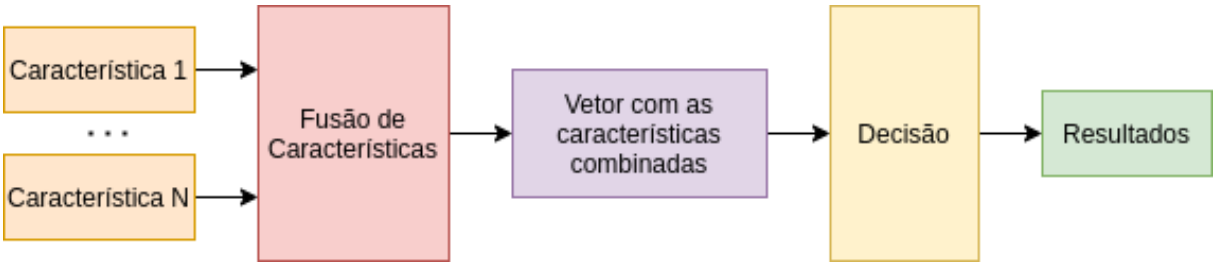
2.3.1 Fusão Prévia

A abordagem de Fusão Prévia possui a objetivo de combinar os vetores de características das informações midiáticas, ou seja, executar um processo de fusão ao nível de características (KISHI, 2020). Inclusive, a vantagem da aplicação desta estratégia em bases de dados heterogêneas é que as análises posteriores serão feitas em cima de um vetor de característica único, além de possibilitar a exploração da correlação entre as informações captadas na etapa inicial no processo de Fusão Multimodal (ATREY *et al.*, 2010).

Segundo Lopes (2015), representar as características das mídias em formatos similares não é uma tarefa simples, pois os dados midiáticos se diferem quanto a suas propriedades. Além disso, de acordo com Lopes (2015) sincronizar diversas características de diferentes mídias é uma atividade computacionalmente complexa e combinar dados de uma maneira precoce onde a dinâmica intermodal é mais complexa, consequentemente propicia a ocorrência de *overfitting* no modelo gerado (BRUIJN *et al.*, 2020).

Por último, na Figura 3 há um exemplo do processo de Fusão Multimodal de modo prévio, onde a princípio diversas características midiáticas são combinadas, posteriormente o vetor resultante com as características combinadas é submetido a um processo de decisão (por exemplo, classificação dos dados através de um algoritmo de AM), assim alcançando um resultado que apoiará a tomada de decisão das pessoas que usufruírem de sistemas embutidos com essa estratégia de Fusão Multimodal.

Figura 3 – Exemplo de Fusão Prévia



Fonte: Adaptada de Atrey *et al.* (2010).

2.3.2 Fusão Tardia

A abordagem de Fusão Tardia segundo Alqhtani, Luo e Regan (2015) possui a tarefa de processar os recursos primários pelas suas respectivas unidades de tomada de decisão e posteriormente encaminhar os resultados combinados para a unidade de análise de dados. Inclusive, esta estratégia de Fusão Multimodal é comumente conhecida na literatura como fusão ao nível de decisão (KISHI, 2020).

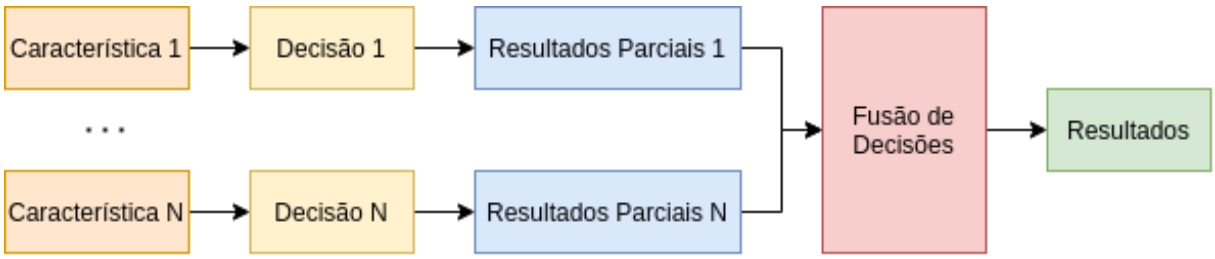
Além disso, segundo Lopes (2015) nesta abordagem de Fusão Multimodal as diversas características das diferentes mídias são processadas independentemente, logo isso proporciona

a aplicação de estratégias mais adequadas (por exemplo, algumas informações midiáticas podem ser combinadas através de regras, já outras por algoritmos de AM), além de que esta abordagem de fusão não exibe problemas de sincronização de características.

Além de que, combinar informações midiáticas de maneira tardia pode propiciar na simplificação excessiva da dinâmica intermodal (BRUIJN *et al.*, 2020). Ademais, de acordo com Lopes (2015) nesta abordagem de Fusão Multimodal a correlação entre os vetores de características não são exploradas, aliás a aplicação de diferentes técnicas para o processamento dos recursos inicias pode demandar um período de processamento maior do que quando comparado com outras abordagens de fusão de dados midiáticos. Inclusive, esta estratégia de Fusão Multimodal tende a ser mais escalável que a Fusão Prévia, visto que há uma simplificação no processo de inserção de novas características ao modelo (LOPES, 2015).

Por último, na Figura 4 há um exemplo do processo de Fusão Multimodal de modo tardio, onde primeiramente as diversas características são submetidas a diferentes processos de decisão, logo após os diversos vetores com os resultados parciais resultantes dos processos de decisão são combinados, assim alcançando resultados que tendem a ser úteis para os profissionais responsáveis pela tomada de decisão em situações críticas.

Figura 4 – Exemplo de Fusão Tardia



Fonte: Adaptada de Atrey *et al.* (2010).

2.3.3 Fusão Híbrida

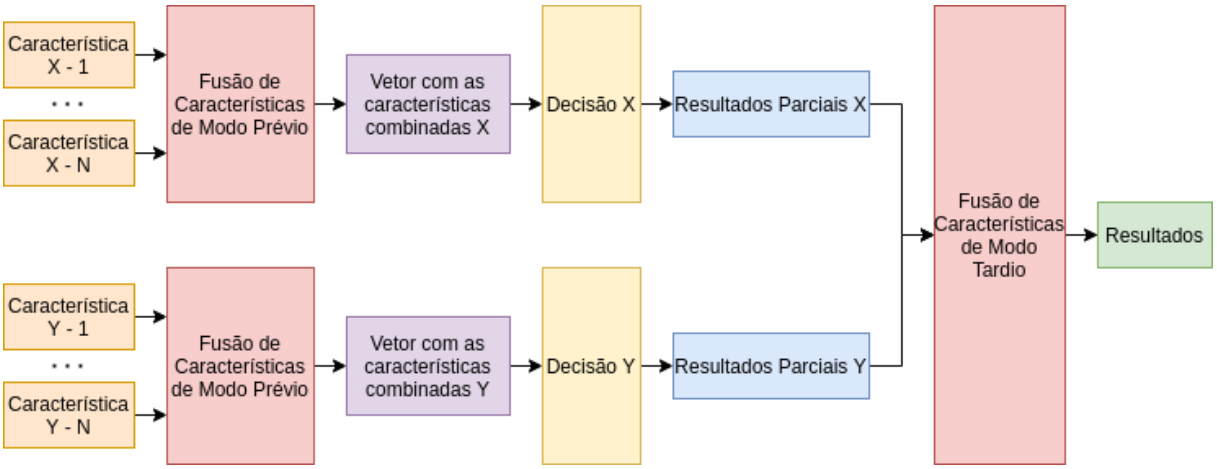
A abordagem de Fusão Híbrida possui a tarefa de combinar de modo prévio uma parcela dos vetores de características das informações midiáticas, posteriormente os vetores parciais resultantes são submetidos ao processo de Fusão Tardia com os demais dados multimodais (LOPES, 2015). Dessa forma, de acordo com Lopes (2015) esta estratégia de fusão combina os benefícios provenientes da Fusão Prévia e da Fusão Tardia. Inclusive, seja qual for a abordagem que combine os processos de Fusão Prévia e Fusão Tardia é considerado como Fusão Híbrida, além de que esta abordagem soluciona os obstáculos apresentados por ambas estratégias de Fusão Multimodal (LOPES, 2015).

Além disso, a estratégia de Fusão Híbrida é aconselhável ser empregada em tarefas que envolvam a utilização de diversas informações de diferentes formatos (por exemplo, dados textuais, informações meteorológicas e dados geográficos) (LOPES, 2015). Desse modo, as

características que são similares podem ser combinadas de modo prévio, já as divergentes podem ser aplicadas ao processo de Fusão Tardia (LOPES, 2015). Outrora, esta abordagem de Fusão Multimodal gera um gasto computacional exacerbado, visto que os vetores resultantes não são obtidos com somente uma camada de decisão, mas com duas ou mais (LOPES, 2015).

Por último, na Figura 5 há um exemplo do processo de Fusão Multimodal de modo híbrido, onde primeiramente as características midiáticas primárias são combinadas de modo prévio e os vetores de características resultantes são submetidos a distintos processos de decisão, assim os diversos resultados parciais resultantes são submetidos ao processo de Fusão Multimodal Tardio e os resultados alcançados podem ser utilizados por profissionais que utilizam sistemas de tomada de decisão em ambientes críticos.

Figura 5 – Exemplo de Fusão Híbrida



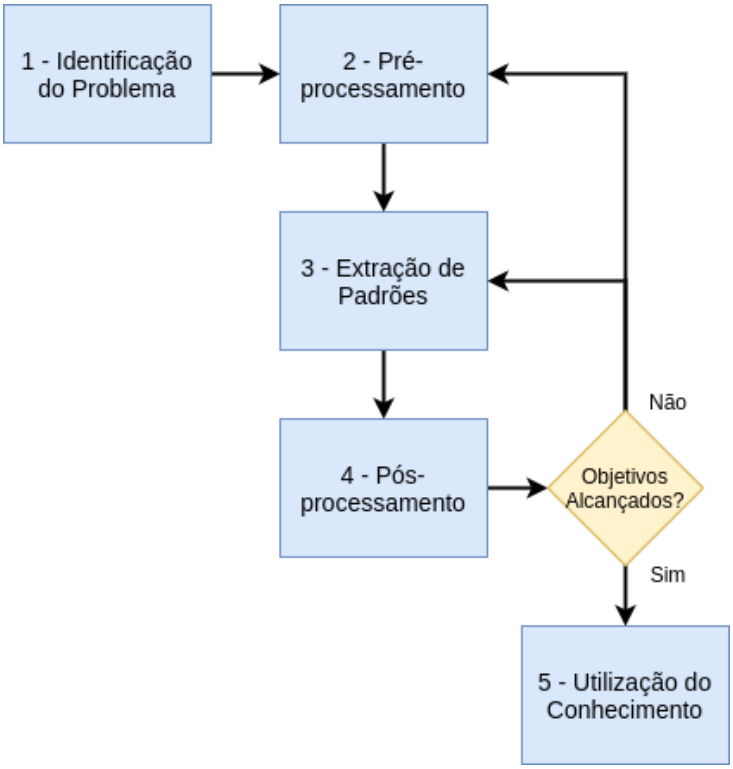
Fonte: Adaptada de Lopes (2015).

2.4 Mineração de Textos

O *Text Mining* (TM), na Língua Portuguesa conhecido como Mineração de Texto (MT), é uma coleção de métodos utilizados para a análise de dados não estruturados com o intuito de proporcionar a descoberta de padrões desconhecidos no espaço amostral estudado (AGGARWAL; ZHAI, 2012). Desse modo, de acordo com Sinoara (2018) a MT pode ser caracterizada como uma especialização da Mineração de Dados (MD), pois a MD é comumente aplicada na literatura para a análise de dados estruturados, já a MT é empregada para a análise do conjunto de informações textuais escritas em linguagem natural (dados não estruturados).

O processo de MT é composto por cinco etapas que podem ser observadas na Figura 6. Desse modo, segundo Sinoara (2018) a primeira fase deste processo é chamada de **Identificação do Problema**, onde possui o objetivo de delimitar o escopo do processo de MT e definir quais serão os textos que farão parte do espaço amostral, além de determinar qual será a finalidade dos resultados alcançados com a aplicação desta metodologia de mineração textual.

Figura 6 – Processo de Mineração de Textos



Fonte: Adaptada de Sinoara (2018).

Já a etapa de **Pré-processamento**, segundo Sinoara (2018) possui o objetivo de preparar os dados não estruturados para a etapa de Extração de Padrões, assim procura-se transformar as informações em formatos apropriados para a captação de conhecimento. Além disso, esta etapa é importante para melhorar a qualidade das informações, pois elas podem exibir diferentes características e dimensões (FACELI *et al.*, 2011). Por exemplo, nos espaços amostrais podem existir observações com dados desconhecidos, ruídos, informações com baixo valor de predição, quantidade de observações de cada classe desproporcionais, entre outros (BATISTA *et al.*, 2003). Desse modo, as estratégias que podem ser utilizadas na etapa de pré-processamento para a solução desses obstáculos são: tratamento de informações desbalanceadas, limpeza de dados, integração de informações, transformações de dados e redução de dimensionalidade (FACELI *et al.*, 2011).

Portanto, para a fase de limpeza da etapa de Pré-processamento da metodologia de MT, é comumente encontrado na literatura algumas pesquisas que aplicam o processo de mineração textual em mensagens captadas de redes sociais as seguintes técnicas de limpeza de dados: filtragem textual, tokenização, remoção das *stopwords* (por exemplo, Salas, Georgakis e Petalas (2017), Aguiar *et al.* (2018), Bruijn *et al.* (2020)). Logo após a fase de limpeza textual, é importante a realização da transformação dos dados, pois diversos algoritmos de AM que serão utilizados na etapa de Extração de Padrões possuem dificuldades para o processamento de informações textuais, então necessita-se empregar estratégias que convertem características

simbólicas para numéricas (por exemplo, BOW, TF-IDF e *Word Embeddings*) (FACELI *et al.*, 2011). Inclusive, *Bag Of Words* (BOW) é uma abordagem de transformação textual, onde cada documento presente no espaço amostral é representado por um vetor de elementos contido no conjunto de dados, ou seja, cada sentença é representada por um vetor com as palavras mais frequentes existentes (MATSUBARA; MARTINS; MONARD, 2003). Aliás, o *Term Frequency-Inverse Document Frequency* (TF-IDF), é uma estratégia que define os pesos das palavras mais frequentes do espaço amostral, onde a relevância delas é dada pelo inverso da frequência da presença desses elementos em todo o conjunto de dados textual (ZHANG; YOSHIDA; TANG, 2011). Neste trabalho, aplicamos as estratégias de tratamento de informações desbalanceadas (por exemplo, o espaço amostral é constituído de 50% de exemplos de cada classe), limpeza de dados (por exemplo, filtragem textual, remoção das *stopwords* e tokenização) e transformação de dados (por exemplo, BOW, TF-IDF, *Word Embeddings* das categorias Word2Vec e Fast Text dos tipos Skip-Gram e CBOW com 50 e 100 dimensões cada).

Posteriormente a etapa de Pré-processamento, os conjuntos de dados textuais resultantes do processo de limpeza de dados, transformação de informações e redução de dimensionalidade são transmitidos para a etapa de Extração de Padrões. Inclusive, na fase de **Extração de Padrões** as atividades executadas são de acordo com a proposta final do processo de captação de conhecimento, além de que os profissionais responsáveis por esta etapa (por exemplo, Cientistas de Dados, Analistas de Dados, entre outros) executam algoritmos de ML no conjunto de informações pré-processadas com o intuito de captar padrões existentes no espaço amostral (SINOARA, 2018). Aliás, caso a proposta do processo de MT seja relacionada com a organização das observações do espaço amostral, então os algoritmos de ML especialistas em classificação e agrupamento podem ser empregados (SINOARA, 2018). Nesta pesquisa, para o processamento das informações textuais foram empregados alguns algoritmos de ML com o aprendizado do tipo supervisionado que são especializados em classificação de dados (por exemplo, SVM, RF, DT, entre outros).

Logo após a etapa de Extração de Padrões, os conhecimentos obtidos a respeito das informações textuais necessitam ser avaliados e compreendidos na fase de Pós-processamento (SINOARA, 2018). Aliás, a etapa de **Pós-processamento** necessita ser orientada pelas propostas definidas no início da execução do processo de MT (SINOARA, 2018). Inclusive, segundo Sinoara (2018) é possível que a avaliação dos resultados captados para esta etapa seja realizada em parceria com um profissional especializado no domínio das informações textuais, por outro lado, esse procedimento pode ser efetuado também por intermédio de estratégias estatísticas de avaliação (por exemplo, acurácia, precisão, entre outras).

Por último, a fase de **Utilização do Conhecimento** é responsável pela disponibilização dos conhecimentos descobertos no espaço amostral para os usuários, caso essas informações encontradas cumpram as propostas definidas no início da metodologia de MT (SINOARA, 2018). Do contrário, outra bateria de testes deve ser executada, modificando os procedimentos

empregados na etapa de Pré-processamento, alterando os parâmetros dos algoritmos de ML utilizados na etapa de Extração de Conhecimento ou reiniciando o processo de MT na fase de Identificação do Problema (SINOARA, 2018).

2.5 Word Embeddings

O Processamento de Linguagem Natural possui a função de processar palavras e documentos textuais, desse modo pesquisadores descobriram a eficácia da representação da informação textual como vetores densos numéricos, pois eles podem ser manipulados matematicamente em operações que vão desde adição até medidas de distância, além de possibilitar a utilização dessas estruturas de dados em diversos algoritmos de Aprendizado de Máquina para a descoberta de conhecimento (ALMEIDA; XEXÉO, 2019).

As técnicas de representação textual baseadas em *Word Embeddings* são amplamente estudadas e empregadas em pesquisas da área de mineração de texto (SINOARA; ANTUNES; REZENDE, 2017), visto que elas proporcionam a representação de cada palavra de um texto em um único vetor capaz de conter toda a informação (COLLOBERT; WESTON, 2008). De acordo com Baroni, Dinu e Kruszewski (2014), os modelos de representação textual que utilizam *Word Embeddings* permitem descobrir a ocorrência das palavras semanticamente. Desse modo, segundo Sundermann *et al.* (2018) existem diversas estratégias de representação de texto utilizando *Word Embeddings*, como: Word2Vec e Fast Text.

O Word2Vec é uma técnica de representação textual em vetores densos numéricos baseado em Aprendizado Profundo, do inglês *Deep Learning* (DL), assim esta abordagem usufrui de um *corpus* textual como entrada e produz vetores densos semânticos como saída (GOOGLE, 2013). Sendo que, ela constrói um vocabulário a partir das palavras presentes no texto e aprende qual é a representação vetorial desse conjunto de dados (GOOGLE, 2013).

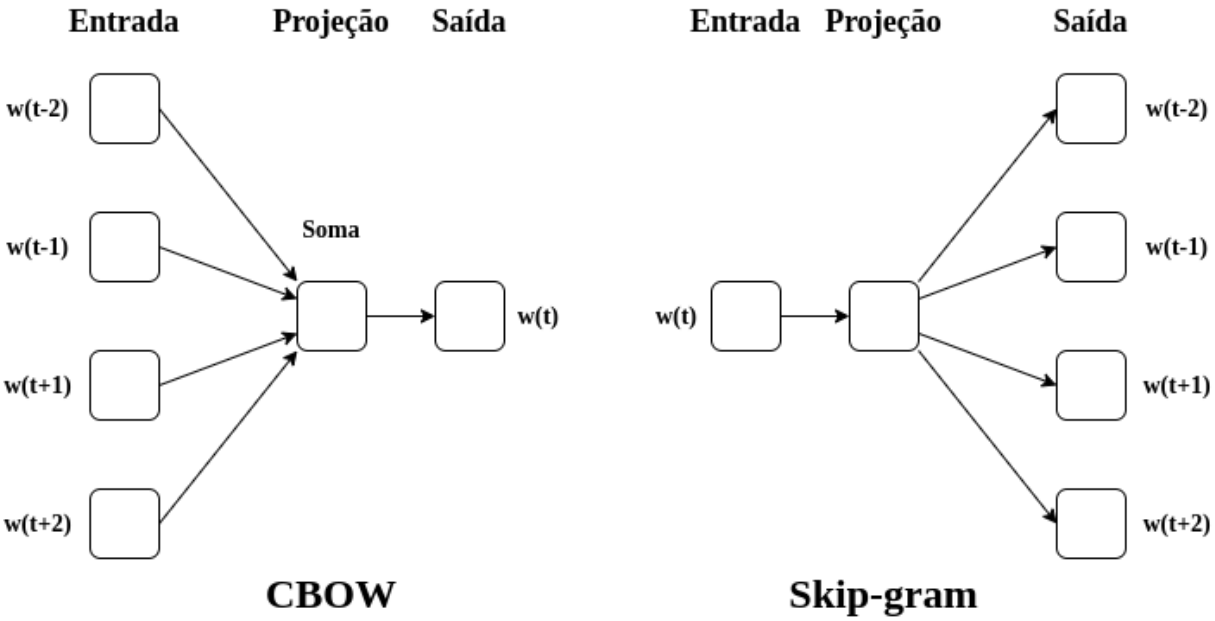
O Fast Text é uma biblioteca de código aberto criada pelos pesquisadores do Facebook[1], no qual proporciona aos usuários estratégias de reconhecimento e representação semântica de palavras em vetores densos numéricos (FACEBOOK, 2020). Além disso, segundo Joulin *et al.* (2016) o Fast Text é uma abordagem de PLN baseada em aprendizado profundo, onde usufrui de uma rede neural multicamadas e produz modelos de classificação de texto precisos e velozes com tamanhos reduzidos.

Os modelos de *Word Embeddings* usufruem de duas arquiteturas comumente empregadas por pesquisadores da área de PLN que são denominados *Continuous Bag of Words* (CBOW) e Skip-Gram (GOOGLE, 2013). Desse modo, os modelos que são baseados em CBOW, a ordem das palavras não influenciará a predição, portanto ele prevê o vetor denso da palavra baseado nas circunstâncias em que ela se encontra (MIKOLOV *et al.*, 2013). Já os modelos que são baseados em Skip-Gram, leva-se em consideração a ordem das palavras para a obtenção da predição, ou

[1] url: <http://facebook.com>

seja, a partir de uma palavra será obtido quais são as similares a ela (MIKOLOV *et al.*, 2013). A seguir na Figura 7, é possível observar a diferença da arquitetura de um modelo baseado em CBOW e outro em Skip-Gram.

Figura 7 – Diferenças entre as arquiteturas de *Word Embeddings* dos tipos CBOW e Skip-gram



Fonte: Adaptada de Mikolov *et al.* (2013).

2.6 Aprendizado de Máquina

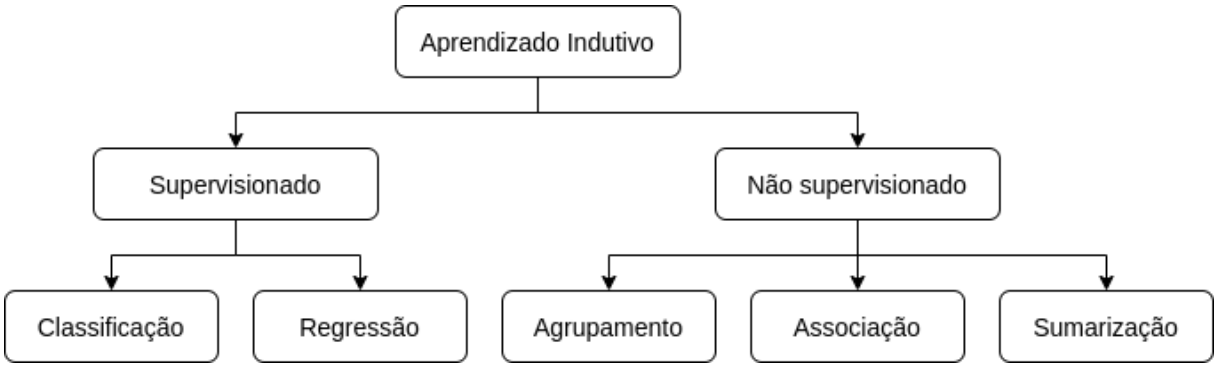
O *Machine Learning* (ML), na Língua Portuguesa é chamado de AM, é um ramo da área da Inteligência Artificial (IA) que possui o objetivo de estudar o reconhecimento de padrões e a teoria do Aprendizado Computacional, dessa forma, este campo de estudo proporciona aos computadores o conhecimento de uma maneira não convencional (SIMON, 2013). Assim, a definição de ML segundo Mitchell (1997), Faceli *et al.* (2011) é:

> "A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência"

Ou seja, os algoritmos de ML aprendem com as experiências que eles foram submetidos, onde esse conhecimento obtido é baseado no princípio da inferência chamada indução, no qual os computadores geram deduções genéricas a partir de uma base de exemplos (FACELI *et al.*, 2011). Desse modo, esses algoritmos geram funções ou suposições capazes de solucionar desafios através de um conjunto de informações que representam instâncias da tarefa que deve ser solucionada (FACELI *et al.*, 2011). Portanto, os algoritmos de ML aprendem com seus erros e fazem predições com os dados que estão sendo analisados (SIMON, 2013).

Os algoritmos de ML podem ser classificados quanto ao paradigma de aprendizado que eles possuem, ou seja, preditivos (aprendizado supervisionado) ou descritivos (aprendizado não supervisionado) (FACELI *et al.*, 2011). Assim, na Figura 8 é apresentado a organização de conhecimento dos algoritmos de ML com base nas diversas categorias de atividades de aprendizado.

Figura 8 – Hierarquia do conhecimento



Fonte: Adaptada de Faceli *et al.* (2011).

Os algoritmos de ML com o aprendizado do tipo supervisionado caracterizam-se pela resolução de tarefas que necessitam da indução de modelos preditivos, assim o conjunto de dados que serão analisados contem atributos de entrada e saída (FACELI *et al.*, 2011). Outrora, os algoritmos de ML que o conhecimento é do tipo não supervisionado são caracterizados pela solução de tarefas que possuem o objetivo de explorar as informações existentes no espaço amostral, assim o conjunto de dados que serão analisados contém somente atributos de entrada (FACELI *et al.*, 2011).

Além disso, na Figura 8 observa-se que há diversos tipos de tarefas de aprendizado que os algoritmos de ML são especialistas, dessa forma as atividades supervisionadas se diferenciam quanto ao rótulo do conjunto de dados analisado, como: discreto para atividades de classificação e contínuo para tarefas de regressão (FACELI *et al.*, 2011). Outrora, de acordo com Faceli *et al.* (2011) os algoritmos que são especialistas em tarefas não supervisionadas se distinguem em: agrupamento, no qual o espaço amostral é separado conforme a similaridade dos dados; associação, onde busca encontrar similaridade entre as características dos dados estudados; sumarização, no qual consistem em descobrir uma definição simples e coesa para o espaço amostral.

Dado o exposto, para alcançar os objetivos deste trabalho foram aplicados algoritmos de ML com o aprendizado do tipo supervisionado (classificação) e não supervisionado (agrupamento) em um conjunto de dados heterogêneo (mensagens de redes sociais, informações meteorológicas e dados geográficos de ocorrências históricas de alagamentos). Inclusive, os algoritmos de ML utilizados nesta pesquisa foram baseados em trabalhos encontrados na literatura cujo objetivo é a obtenção da SAW de eventos a partir das mensagens publicadas em

redes sociais (por exemplo, Sakaki, Okazaki e Matsuo (2010), Yin *et al.* (2012), Boettcher e Lee (2012), Huang e Xiao (2015), Salas, Georgakis e Petalas (2017), Feng e Sester (2018), Andrade *et al.* (2018), Bruijn *et al.* (2020)). Portanto, a seguir serão descritos sucintamente os algoritmos de ML utilizados nesta pesquisa de mestrado. Inclusive, todos esses algoritmos estão disponíveis na biblioteca Scikit-Learn, no qual foi escrita na Linguagem de Programação Python (PEDREGOSA *et al.*, 2011).

2.6.1 Random Forest

O RF é um algoritmo de ML com o aprendizado do tipo supervisionado, onde existe a construção de diversas árvores de decisão para solucionar problemas de classificação de dados (HO, 1995). As árvores de decisão funcionam como fluxogramas, no qual cada nó aponta um teste que foi executado conforme uma condição pré-estabelecida (HO, 1995). Desse modo, as conexões entre os nós das árvores significam os possíveis valores resultantes dos nós superiores, já as folhas apontam a qual classe a informação é pertencente (HO, 1995). Inclusive, segundo Crepaldi *et al.* (2011) o RF usufrui da estratégia de separação para a conquista, isto significa, um desafio complexo é dividido em diversos problemas, assim aplica-se este algoritmo recursivamente para cada sub-problema criado.

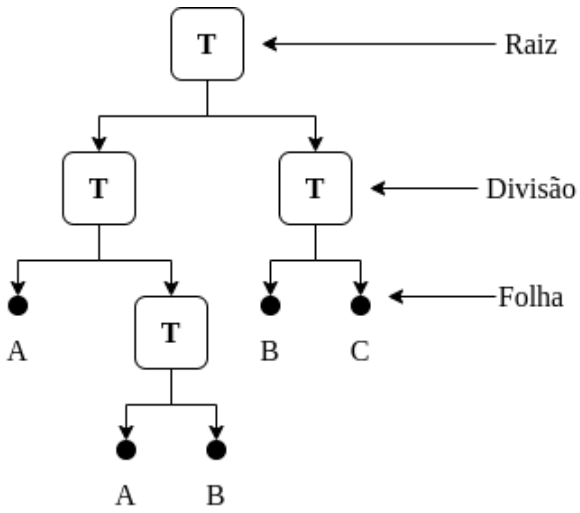
2.6.2 Decision Tree

O DT, na Língua Portuguesa é chamado de Árvore de Decisão (AD), é um algoritmo de ML com o aprendizado do tipo supervisionado, no qual particiona recursivamente o conjunto de dados analisado em subdivisões menores conforme os testes estabelecidos em cada "nó" da árvore (FRIEDL; BRODLEY, 1997). Inclusive, segundo Faceli *et al.* (2011) este algoritmo de ML para a resolução dos problemas em que ele é empregado usufrui da estratégia de separação para a conquista, em outros termos, um desafio que é considerado complexo é separado em diversos problemas simples recursivamente conforme uma estratégia pré-definida.

A seguir na Figura 9, há um exemplo de árvore criada por um algoritmo de DT, onde cada caixa com um "T" inserido corresponde a um teste aplicável e as categorias "A", "B", "C" nos "nós" folhas são as classes resultantes dos experimentos realizados (FRIEDL; BRODLEY, 1997).

No algoritmo DT a árvore gerada é composta por um "nó" raiz, no qual é criado a partir de todos os dados do espaço amostral analisado (FRIEDL; BRODLEY, 1997). Além disso, há um conjunto de "nós" responsáveis pelas divisões, onde inclui testes condicionais baseados nos valores das características dos dados (FACELI *et al.*, 2011). Ademais, de acordo com Friedl e Brodley (1997) existe um grupo de "nós" terminais que são denominados folhas que contém funções. Neste algoritmo de ML, cada "nó" possui somente um "nó" pai e os demais "nós" são filhos ou descendentes (FRIEDL; BRODLEY, 1997). Por último, a classificação realizada pelo

Figura 9 – Exemplo de árvore gerada por um algoritmo de DT



Fonte: Adaptada de Friedl e Brodley (1997).

DT é feita conforme a subdivisão criada pela árvore, onde é dado uma categoria para cada teste conforme o "nó" folha que ele resultar (FRIEDL; BRODLEY, 1997).

2.6.3 Naive Bayes

O NB é um algoritmo de ML baseado no Teorema de Thomas Bayes, no qual segue o paradigma da probabilidade, além de ser um algoritmo cujo aprendizado é do tipo supervisionado (FACELI *et al.*, 2011). Inclusive, neste teorema que derivou o algoritmo NB, os autores descrevem que a probabilidade da ocorrência de um fenômeno B, é estimada pela investigação da frequência de acontecimento deste fenômeno (FACELI *et al.*, 2011). Além disso, segundo Faceli *et al.* (2011), o Teorema de Bayes explica que a probabilidade de um fenômeno ou objeto ser pertencente a uma classe (P(A|B)) é obtido a partir dos seguintes cálculos: a probabilidade a priori da categoria (P(A)); a probabilidade de investigação de vários objetos pertencentes a categoria (P(B|A)); por último, a probabilidade da ocorrência desses objetos (P(B)).

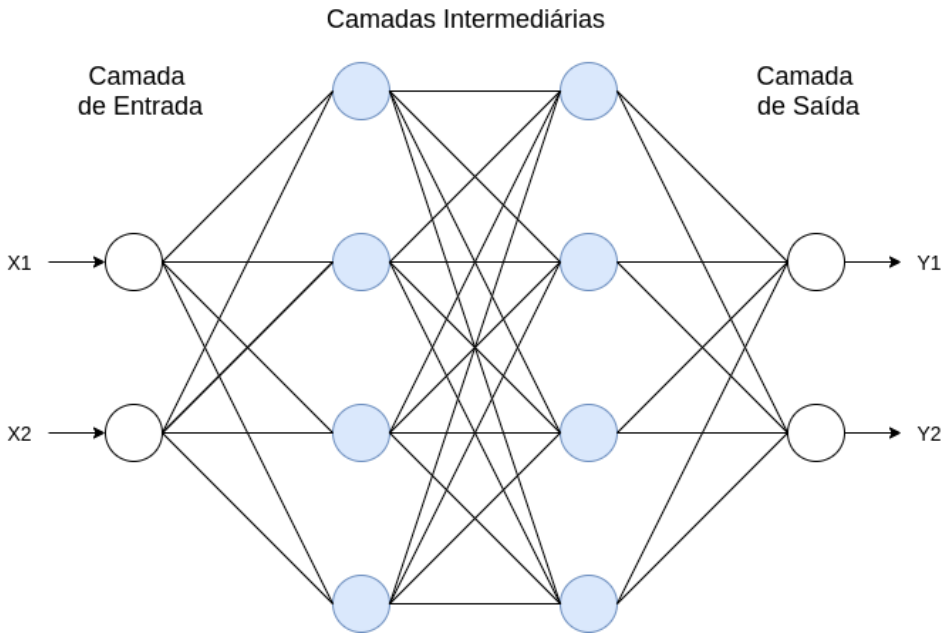
Além disso, um classificador NB é um exemplo específico de Rede Bayesiana, sendo de fácil implementação e de forma incremental, além de que em problemas que os rótulos dos dados são definidos por atributos booleanos a superfície de decisão será linear (FACELI *et al.*, 2011). Ademais, conforme Domingos e Pazzani (1997), o NB tem uma exímia atuação em diversos domínios, além de ter um bom comportamento na presença de dados com atributos irrelevantes e ruídos (KONONENKO, 1991).

2.6.4 Multilayer Perceptron

O *Multilayer Perceptron* (MLP) é um algoritmo de ML do tipo Redes Neurais Artificiais (RNA) com o aprendizado do tipo supervisionado, no qual é aplicado para resolver desafios

de classificação e regressão de dados (BRAGA, 2007). Dessa forma, a MLP é constituída de camada de entrada, camadas intermediárias e camada de saída, como observado na Figura 10.

Figura 10 – Exemplo de MLP



Fonte: Adaptada de Faceli *et al.* (2011).

Os neurônios das camadas de entrada contribuem com as retas que formarão a superfície do espaço de entrada, já os neurônios da camada intermediária combinam as retas dos neurônios da camada anterior formando regiões convexas (BRAGA, 2007). Por último, os neurônios da camada de saída criam áreas que são combinações de regiões convexas (BRAGA, 2007). Dessa forma, é possível categorizar elementos quanto a presença na superfície do espaço (BRAGA, 2007).

Além disso, de acordo com (BRAGA, 2007) para o treinamento deste algoritmo, há a etapa de *feed forward* e *back propagation*, sendo que na primeira etapa a informação é propagada da primeira camada até a última e processada em cada neurônio por uma função de ativação. Já na segunda etapa, os pesos são modificados conforme a regra delta generalizada, logo a informação parte das últimas camadas para as primeiras, onde esse procedimento é repetido até alcançar determinados critérios de parada e consequentemente alcançar a redução do erro encontrados (BRAGA, 2007).

2.6.5 Support Vector Machine

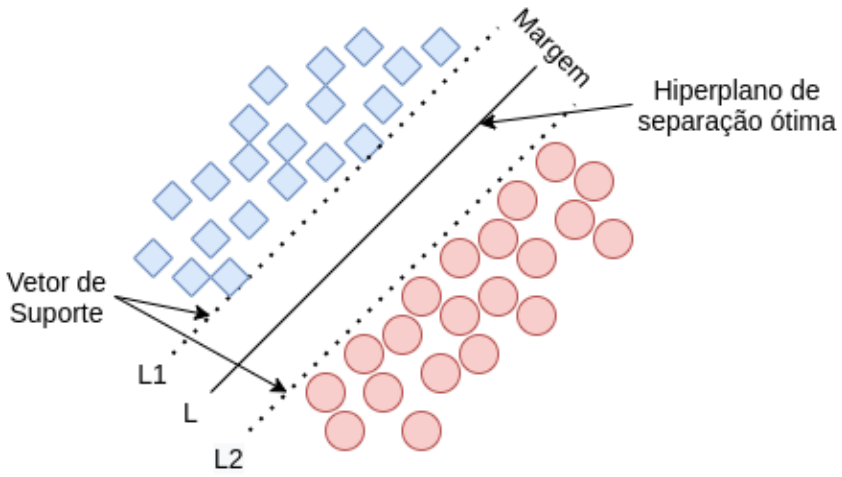
O SVM é um algoritmo de ML que foi criado baseado na teoria do aprendizado estatístico, além de que seu aprendizado é do tipo supervisionado (FACELI *et al.*, 2011). Inclusive, o SVM é aplicado em tarefas de classificação e regressão de dados (FACELI *et al.*, 2011). O propósito deste

algoritmo é construir hiperplanos como superfícies de decisão, de forma que seja perceptível a separação das classes criadas pelo algoritmo (FACELI *et al.*, 2011).

Ademais, segundo Lorena e Carvalho (2007) há duas categorias de SVM: linear e não linear. O primeiro é aplicado quando a base de dados é linearmente separável, desse modo, caso para a execução da tarefa desejada as amostras não possam ser linearmente separáveis, logo é aconselhável a utilização de SVM não linear (LORENA; CARVALHO, 2007). Aliás, caso a aplicação tenha uma base de dados que não seja linearmente separável, portanto é aconselhável a exploração de diferentes categorias de *kernels* (por exemplo, Polinomiais, Gaussianos, Sigmoidais, entre outros), pois eles aumentam a dimensão do conjunto amostral e permite que os dados possam ser divididos por hiperplanos, assim eleva-se o grau de abstração (LORENA; CARVALHO, 2007).

Dado o exposto, a seguir na Figura 11 observa-se um exemplo do processo de classificação efetuado por um SVM, sendo que a separação exímia entre as classes efetuada por um SVM acontece por intermédio de um hiperplano condicional (*L*), no qual é conduzido para potencializar as fronteiras criadas no espaço amostral (distâncias entre as margens *L1* e *L2*) com o intuito de que a divisão das categorias seja mais compreensiva (NASCIMENTO *et al.*, 2009).

Figura 11 – Exemplo de um processo de classificação executado pelo SVM



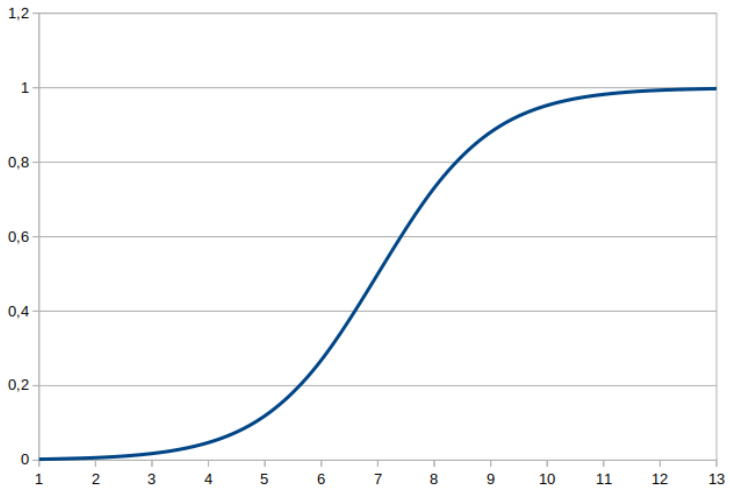
Fonte: Adaptada de Huang, Davis e Townshend (2002), Melgani e Bruzzone (2004), Nascimento *et al.* (2009).

2.6.6 Logistic Regression

O *Logistic Regression* (LR), na Língua Portuguesa é chamado de Regressão Logística (RL), é um algoritmo de ML com o aprendizado do tipo supervisionado que é utilizado para o cálculo ou predição de acontecimentos de fenômenos específicos (FIGUEIRA, 2006). Inclusive, o algoritmo de LR pode ser do tipo ordinal ou nominal dependendo do tipo da natureza dos valores das variáveis dependentes (FIGUEIRA, 2006). Desse modo, a relação entre as características consideradas dependentes e a independente quando visualizadas em editores de planilhas

assemelha-se a uma curva com o formato de "S" (curva sigmoide) (GONÇALVES; GOUVÊA; MANTOVANI, 2013). A seguir na Figura 12, há um exemplo de uma curva sigmoide.

Figura 12 – Exemplo de uma curva gerada por uma função logística



Fonte: Elaborada pelo autor.

Além disso, para o entendimento da equação matemática que originou o algoritmo classificador LR, será dado o seguinte exemplo de análise de risco financeiro, a princípio devemos observar um caso de classificação binária de observações (0 ou 1), onde representam uma pessoa que se caracteriza como um bom pagador ou um mau pagador perante a um analista de riscos financeiros (GONÇALVES; GOUVÊA; MANTOVANI, 2013). Desse modo, pode ser atribuído os valores a característica dependente binária "Y": "1", caso a i-ésima pessoa classificada seja pertencente a classe dos bons pagadores, outrora "0", caso o i-ésimo cliente rotulado pertença à categoria dos maus pagadores (GONÇALVES; GOUVÊA; MANTOVANI, 2013).

O modelo de regressão logística é uma ocorrência específica dos Modelos Lineares Generalizados (PAULA, 2004; GONÇALVES; GOUVÊA; MANTOVANI, 2013; DOBSON; BARNETT, 2018). Assim, dado X = $(X_1, X_2, ..., X_n)$, no qual corresponde a um vetor que o elemento primário equivale a 1 (constante) e os outros são equivalentes às diversas características do modelo (GONÇALVES; GOUVÊA; MANTOVANI, 2013). Portanto, a função matemática que define este modelo é:

$$ln\left(\frac{p(X)}{1-p(X)}\right) = \beta' X = Z \tag{2.1}$$

Dessa forma, de acordo com Gonçalves, Gouvêa e Mantovani (2013), $\beta' = (\beta_1, \beta_2, ..., \beta_n)$ é equivalente ao vetor das características associadas as variáveis, além disso, a probabilidade de uma pessoa ser rotulada como um bom ou mau pagador é P(X) = E (Y=1|X) (GONÇALVES;

GOUVÊA; MANTOVANI, 2013). Sendo que, essa probabilidade é definida da seguinte maneira por Neter *et al.* (1996):

$$p(X) = E(Y) = \frac{e^{\beta' X}}{1 + e^{\beta' X}} = \frac{e^{Z}}{1 + e^{Z}} \tag{2.2}$$

2.6.7 Density-Based Spatial Clustering of Applications with Noise

O *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), é um algoritmo de agrupamento que não determina previamente a quantidade de *clusters* que serão formados, possui uma exímia competência para o processamento de grandes volumes de informações e tem a habilidade de identificar grupos de diferentes formas (BIRANT; KUT, 2007). Além disso, este algoritmo possui dois parâmetros essenciais: a quantidade mínima de grupos e o raio dos agrupamentos. A quantidade mínima de grupos é definida como 1, já o raio dos agrupamentos é a razão de um limiar pelo raio equatorial da terra (BIRANT; KUT, 2007).

Ademais, o DBSCAN é um algoritmo de agrupamento baseado na densidade das informações, ou seja, ele tenta reconhecer grupos pela análise da densidade dos pontos distribuídos no espaço, visto que dentro de cada grupo a densidade é maior que fora (BORAH; BHATTACHARYYA, 2004). Aliás, este algoritmo de clusterização de alto desempenho consegue encontrar agrupamentos de maneira arbitrária lidando de modo eficaz com os ruídos (BORAH; BHATTACHARYYA, 2004).

Segundo Borah e Bhattacharyya (2004), o algoritmo DBSCAN cria agrupamentos através da busca nos vizinhos de cada ponto no espaço, dessa forma caso na vizinhança haja mais pontos que o limite mínimo de pontos pré-determinado para a formação de grupos e a distância entre eles sejam inferior ao raio de formação de agrupamentos, então um novo *cluster* será formado. Assim, o DBSCAN reúne iterativamente objetos acessíveis conforme a densidade local, portanto pode ocorrer a junção de alguns grupos (BORAH; BHATTACHARYYA, 2004). Por último, o processo de agrupamento é finalizado quando nenhum ponto no espaço pode ser inserido em nenhum outro grupo (BORAH; BHATTACHARYYA, 2004).

2.6.8 Ordering Points To Identify the Clustering Structure

O *Ordering Points To Identify the Clustering Structure* (OPTICS), é um algoritmo de ML com o aprendizado do tipo não supervisionado e que não define uma quantidade máxima de agrupamentos que serão criados, além de proporcionar uma ordenação ampliada dos objetos da base de dados que estão sendo analisados (ANKERST *et al.*, 1999). Ademais, esta abordagem de agrupamento é baseada em densidade, desse modo segundo Ankerst *et al.* (1999) a principal característica desse tipo de algoritmo não supervisionado é que para cada observação do espaço amostral agrupado, a sua vizinhança deve conter uma quantidade mínima de objetos, ou seja, a cardinalidade da vizinhança dos agrupamentos deve exceder um limiar pré-definido.

Além disso, o OPTICS proporciona uma exploração interativa da estrutura dos grupos criados e oferece para o usuário a oportunidade de compreender sobre a distribuição e a correlação dos dados analisados (ANKERST *et al.*, 1999). Inclusive, de acordo com Ankerst *et al.* (1999) quando comparado este algoritmo com as demais abordagens de agrupamento presentes na literatura nota-se que este algoritmo não necessita da definição de parâmetros globais.

Por último, segundo (SCIKIT-LEARN, 2020b) o OPTICS quando comparado a outras estratégias de agrupamento baseadas em densidade (por exemplo, DBSCAN) mantém a hierarquia dos grupos criados para um raio de vizinhança flexível, além de que esta estratégia é mais pertinente para a aplicação em grandes conjuntos de dados.

2.6.9 Hierarchical Agglomerative Clustering

O *Hierarchical Agglomerative Clustering* é um algoritmo de ML com o aprendizado do tipo não supervisionado, no qual está contido no conjunto de algoritmos de ML com a categoria hierárquica, inclusive esta categoria de algoritmo possui a característica de gerar uma sequência de subdivisões aninhadas a partir de uma matriz de adjacências (FACELI *et al.*, 2011). Além disso, os algoritmos hierárquicos podem ser classificados como: **aglomerativos** ou **divisivos**. Desse modo, os **algoritmos hierárquicos aglomerativos** possuem a característica de iniciar a execução com n agrupamentos num único objeto e sequencialmente formam as diversas subdivisões agrupando os *clusters* consecutivamente (FACELI *et al.*, 2011). Já os **algoritmos hierárquicos divisivos**, o processo de clusterização é iniciado com um *cluster* com todos os objetos e a sequência é construída a partir da divisão sucessiva dos agrupamentos (FACELI *et al.*, 2011).

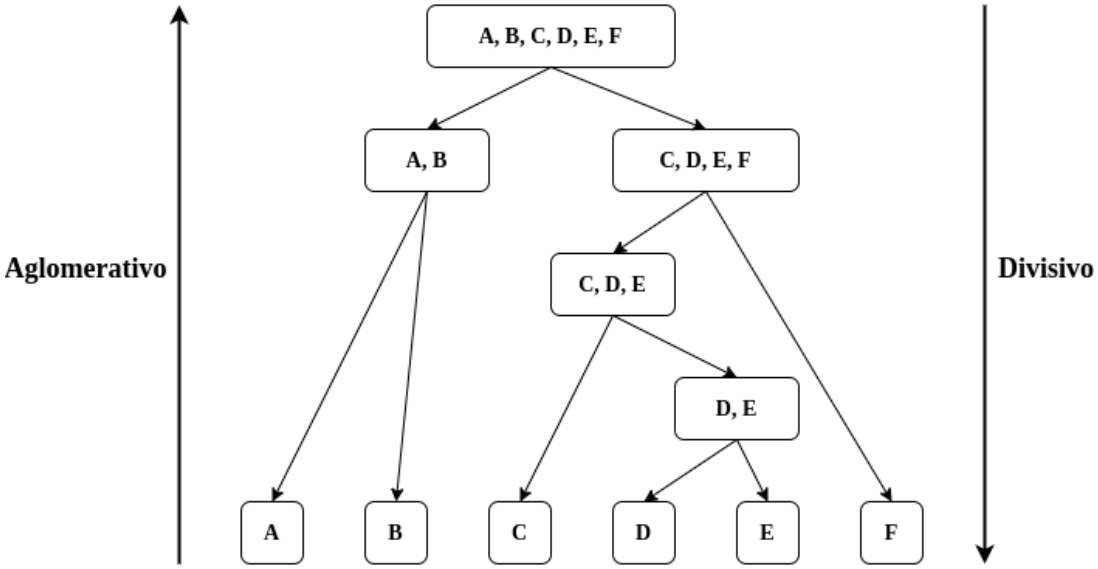
Ademais, segundo Faceli *et al.* (2011) os algoritmos hierárquicos possuem algumas vantagens quando comparado a outros algoritmos do tipo não supervisionado, como: versatilidade quanto ao grau de granularidade das tarefas analisadas, facilidade na aplicação de diversos modos de similaridade ou distância e a oportunidade de empregar indeterminadas categorias de atributos. Outrora, as desvantagens são: não há a melhora dos agrupamentos após a finalização dos processos de criação (FACELI *et al.*, 2011).

As abordagens de clusterização do tipo hierárquico usufruem de métricas de integração (*linkage metrics*), no qual são medidas de distância entre os agrupamentos que funcionam como critério para a união ou divisão de *clusters* (FACELI *et al.*, 2011). Dessa forma, segundo Müllner (2011) os tipos de ligação mais comumente encontrados na literatura para a aplicação nesses algoritmos são: ***single***, ***complete***, ***average*** e ***ward***. Inclusive, o critério de ligação do tipo ***single*** caracteriza-se pela utilização da distância mínima entre as observações para a formação de grupos (SCIKIT-LEARN, 2020a). Já o critério de ligação classificado como ***average***, caracteriza-se pela aplicação da distância média das observações como critério para a formação de agrupamentos (SCIKIT-LEARN, 2020a). Outrora, o modo de ligação do tipo ***complete***, é caracterizado pela utilização das distâncias máximas das observações do conjunto de dados analisado para a forma-

ção dos *clusters* (SCIKIT-LEARN, 2020a). Por último, o modo de ligação classificado como ***ward***, caracteriza-se pela minimização das variâncias dos agrupamentos que são combinados (SCIKIT-LEARN, 2020a).

A seguir na Figura 13, há um exemplo do funcionamento dos algoritmos hierárquicos classificados como aglomerativos e divisivos.

Figura 13 – Exemplo das atividades dos algoritmos hierárquicos do tipo aglomerativo e divisivo



Fonte: Adaptada de Faceli *et al.* (2011).

Dado o exposto, observa-se na Figura 13 os processos de agregação e desagregação dos grupos nos algoritmos hierárquicos aglomerativos e divisivos (FACELI *et al.*, 2011). Dessa forma, com o intuito de agrupar ou dividir os *clusters* os algoritmos utilizam estratégias de cálculo de distância, ou similaridade definidas pelas diversas classes de critérios de ligação (por exemplo, *single*, *complete*, *average* e *ward*), sendo que cada modo de integração influencia diretamente a execução dos algoritmos hierárquicos (FACELI *et al.*, 2011). Inclusive, esses algoritmos não possuem funcionalidades globais, ou seja suas decisões são realizadas somente localmente (FACELI *et al.*, 2011).

Conclui-se que, nos algoritmos hierárquicos aglomerativos a metodologia de formação de *clusters* baseia-se no agrupamento de observações mais próximas a partir de um modo de ligação pré-definido, outrora os algoritmos hierárquicos categorizados como divisivos possuem o objetivo de dividir os *clusters* com o intuito de gerar subdivisões mais distintas a partir de uma métrica de integração pré-definida (FACELI *et al.*, 2011).

2.7 Geoestatística

O surgimento da Geoestatística é datado de 1951 na África do Sul, país pertencente ao continente Africano, pelo engenheiro Daniel Krige e o estatístico H. S. Sichel quando eles trabalhavam com um conjunto de dados de acúmulo de ouro e necessitavam descobrir a localização das reservas, assim os pesquisadores descobriram de forma empírica uma técnica capaz de estimar o cálculo de reservas minerais (KRIGE, 1951; LANDIM, 2006).

Desse modo, os pesquisadores ao trabalharem com o conjunto de dados de reservas de minério perceberam que apenas as informações das variâncias eram incompletas para descrever as ocorrências das jazidas de ouro, logo era imprescindível considerar a distância entre as observações do espaço amostral (KRIGE, 1951; LANDIM, 2006). Assim, G. Matheron elaborou a teoria das variáveis regionalizadas baseando-se nas observações realizadas por Daniel Krige, onde uma variável regionalizada é uma aplicação numérica com distribuição espacial (MATHERON, 1963; LANDIM, 2006).

Dessa forma, os pesquisadores Daniel Krige e G. Matheron são os responsáveis pelo desenvolvimento da "Geoestatística", no qual é uma subárea da estatística que usufrui da concepção de variáveis regionalizadas para a resolução de desafios de mutabilidade espacial, inclusive a "Geoestatística" é uma imensa contribuição da Geologia para a Estatística Prática (LANDIM, 2006).

Além disso, de acordo com Isaaks e Srivastava (1989) a área da Geoestatística não é limitada a somente a obtenção de um modelo de dependência espacial, visto que propicia a inferência da quantidade de minérios de locais que não estão contidos no espaço amostral. Ou seja, a Geoestatística proporciona estimar valores de uma propriedade pertencente um local que não foi medido, a partir da aplicação de uma função de correlação espacial entre as informações sem vieses e com variâncias irrelevantes (VIEIRA *et al.*, 2000).

Portanto, ao comparamos a estatística clássica com a Geoestatística, nota-se que a primeira precisa de normalidade e autonomia espacial entre as informações do espaço amostral analisado, já a segunda é preciso que as informações tenham auto-correlação espacial, pois há a possibilidade da organização dos dados do espaço amostral conforme a semelhança entre elas (ISAAKS; SRIVASTAVA, 1989). Inclusive, segundo Landim (2006) a partir dos cálculos disponíveis na área da Geoestatística é impossível inferir com exatidão os teores dos minérios presentes na jazida, no entanto, é provável que haja reservas de minérios próximas de locais com reservas em abundância.

Conclui-se que esta subárea da estatística é muito importante para os estudos que necessitam da obtenção de um modelo de dependência espacial, pois a partir de dados auto-correlacionáveis é possível encontrar determinados pontos de interesse em regiões inexploradas.

2.7.1 Semivariograma

O Semivariograma é uma técnica explorada na área da Geoestatística, no qual tem o intuito de verificar a presença de dependência espacial entre as observações de uma amostragem georreferenciada, sendo que ele pode ser representado por um gráfico da função de semivariância pela sua distância (ISAAKS; SRIVASTAVA, 1989). Além disso, a distância máxima que uma função de semivariância é estimada, chama-se *Cut-Off* e os pontos que se encontram estabelecidos após o *Cut-Off* são desconsiderados (ISAAKS; SRIVASTAVA, 1989).

Ademais, o Semivariograma é uma técnica da Geoestatística que mostra o nível de dependência espacial entre as observações de um espaço amostral em um suporte exclusivo (LANDIM, 2006). Aliás, esta técnica mede a mutabilidade geológica vinculada a distância geográfica, ou seja, a mutabilidade se diferencia quando considerada em direções distintas (LANDIM, 2006).

Para a construção de um Semivariograma os valores do espaço amostral devem estar contidos em um intervalo regular (LANDIM, 2006). Além disso, segundo Landim (2006) é imprescindível a utilização de uma função de semivariância, sendo que o resultado dessa função deve ser ajustado a um modelo teórico. Inclusive, a semivariância é uma técnica para a análise da variabilidade espacial do fenômeno estudado (FÉLIX *et al.*, 2016). Dessa forma, a seguir encontra-se a fórmula da semivariância:

$$\gamma(h) = \frac{1}{2n}\sum(x_{i+h} - x_i)^2 \tag{2.3}$$

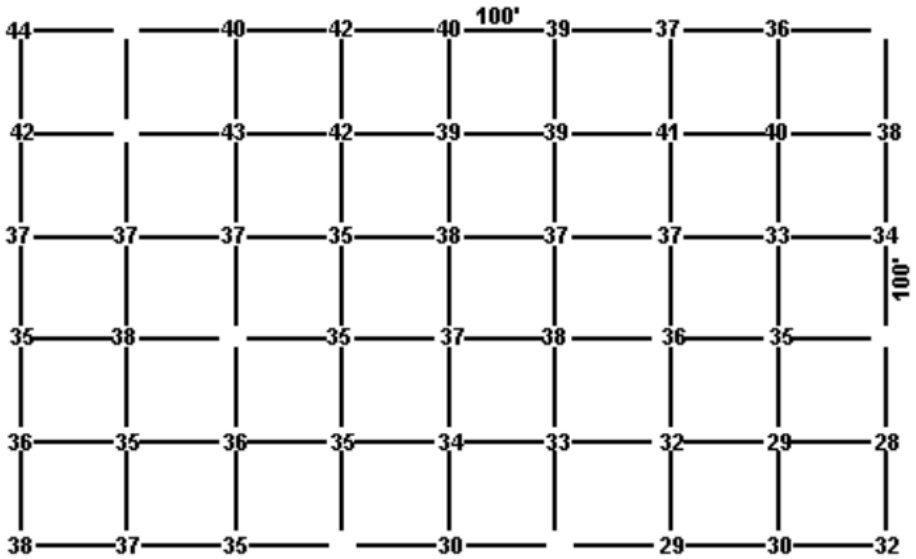
Dado $(X_1, X_2, ..., X_n)$ os valores de uma variável regionalizada, observa-se na Equação 2.3, o cálculo do Semivariograma em uma determinada direção (h), no qual esta equação corresponde a somatória dos quadrados das diferenças entre o valor de cada amostra geolocalizada dentro de um limiar de distância e posteriormente dividida por duas vezes a quantidade dessas diferenças (CLARK, 1979).

No trabalho de Clark (1979), é calculado o Semivariograma de Leste para oeste e norte para sul de uma rede regular de depósitos de ferro, onde cada amostra tem 100 pés (30,48 metros) de distância. Desse modo, na Figura 14 nota-se o mapa com a localização das observações do espaço amostral de minérios de ferro.

Dado o exposto, para que seja possível calcular o Semivariograma da Figura 14 é necessário escolher um valor inicial, qual direção será analisada e definir um limiar de distância para o estudo (LANDIM, 2006). Desse modo, os resultados obtidos no cálculo da função de Semivariograma para as orientações leste ao oeste, podem ser observadas na Tabela 1, já os resultados dessa função para as direções norte ao sul, nota-se na Tabela 2.

Além disso, observa-se na Figura 15 que os resultados alcançados são mais contínuos na direção leste para oeste do que na norte para sul. Aliás, a partir desses resultados é possível ajustá-

Figura 14 – Mapa com a localização das reservas de minério de ferro



Fonte: Clark (1979).

Tabela 1 – Resultados da aplicação da função de Semivariograma para orientações leste ao oeste

Distância	**Semivariograma**	**Número de Pares**
100	1,46	36
200	3,30	33
300	4,31	27
400	6,70	23

Fonte: Adaptada de Clark (1979).

Tabela 2 – Resultados da aplicação da função de Semivariograma para orientações norte ao sul

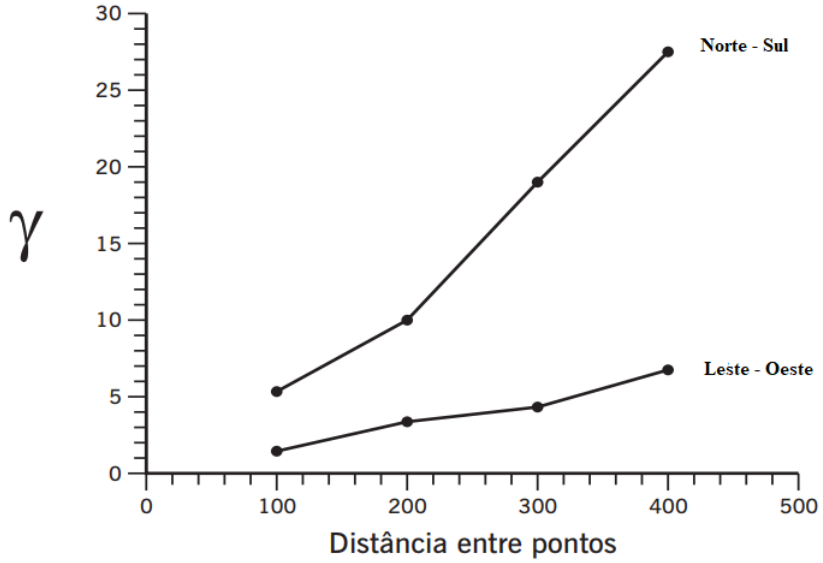
Distância	**Semivariograma**	**Número de Pares**
100	5,35	36
200	9,87	27
300	18,88	21

Fonte: Adaptada de Clark (1979).

los a modelos teóricos com o intuito da obtenção de *insights* sobre o modelo de dependência espacial (por exemplo, a obtenção da distância que as observações apresentam correlação espacial).

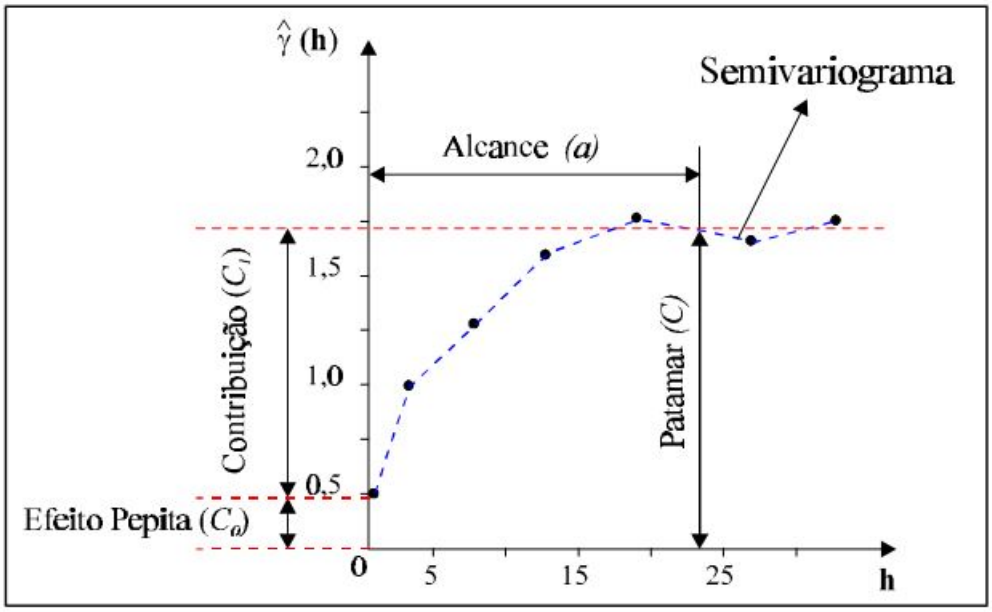
Ademais, na Figura 16, encontra-se um exemplo de um semivariograma experimental, no qual observa-se os parâmetros: **efeito pepita** (C_0), **contribuição** (C_1), **alcance** (α) e **patamar** $(\boldsymbol{C})$.

Figura 15 – Resultados dos Semivariogramas contabilizados para as orientações leste ao oeste e norte ao sul



Fonte: Landim (2006).

Figura 16 – Exemplo de Semivariograma Experimental



Fonte: Camargo (1998).

A seguir, os parâmetros observados na Figura 16 são descritos:

- **Efeito Pepita** (C_0): O Efeito Pepita informa a interrupção do Semivariograma para distâncias inferiores ao limiar das menores distâncias contidas nas observações do espaço amostral (ISAAKS; SRIVASTAVA, 1989).
- **Contribuição** (C_1): A Contribuição é definida como o diferencial alcançado entre o

Patamar (C) e o Efeito Pepita (C_0) (CAMARGO, 1998).

- **Alcance** (α): O Alcance é a distância máxima das observações do espaço amostral que apresentam correlação espacial (CAMARGO, 1998).

- **Patamar** $(\boldsymbol{C})$: O valor do Patamar é equivalente ao resultado da estabilização da função da semivariância $\gamma(h)$ (ISAAKS; SRIVASTAVA, 1989). Desse modo, de acordo com Isaaks e Srivastava (1989), o valor de $\gamma(h)$ aumenta proporcionalmente conforme o crescimento de h até que atinja um valor máximo de estabilização, ou seja, quando h é equivalente ao alcance (α).

CAPÍTULO

3

TRABALHOS RELACIONADOS

Considerando que esta dissertação elabora uma abordagem de Fusão Multimodal capaz de obter SAW de alagamentos a partir de mensagens publicadas no Twitter com o intuito de auxiliar a etapa de resposta da GD, então neste capítulo na Seção 3.1 serão apresentados alguns trabalhos correlatos a esta pesquisa que foram publicados em conferências e periódicos internacionais, posteriormente na Seção 3.2 é feita uma síntese dos trabalhos apresentados, além de uma comparação entre as pesquisas correlatas e a abordagem desenvolvida nesta dissertação. Por último, na Seção 3.3 são apresentadas as considerações finais deste capítulo.

3.1 Discussão dos Trabalhos Relacionados

De acordo com uma pesquisa realizada em repositórios acadêmicos das áreas de Gestão de Desastres, Fusão de Dados, Mineração de Texto, Análise de Redes Sociais, Aprendizado de Máquina, foram escolhidos os seguintes trabalhos como correlatos por possuírem características semelhantes ao do trabalho elaborado nesta dissertação:

- ***Erthquake shakes twitter users: real-time event detection by social sensor*** (SAKAKI; OKAZAKI; MATSUO, 2010).

  Nesse artigo os autores realizaram o desenvolvimento de uma abordagem computacional para o monitoramento de terremotos e tufões no Japão que usufrui de mensagens publicadas no *Twitter*, onde os *tweets* relacionados aos fenômenos são categorizados como positivos, já as demais mensagens são rotuladas como negativas (SAKAKI; OKAZAKI; MATSUO, 2010). Inclusive, os autores desse trabalho desenvolveram um modelo de classificação de *tweets* com o intuito de categorizar as informações relativas a terremotos e tufões, assim a partir desses dados foi possível a identificação da trajetória geográfica dos fenômenos naturais e a notificação da população residente das regiões afetadas via *software* e *e-mail* (SAKAKI; OKAZAKI; MATSUO, 2010).

Os autores dessa pesquisa informam que para a criação do classificador de *tweets* eles treinaram um SVM com *kernel* do tipo linear baseado num espaço amostral de palavras-chave relacionadas com desastres naturais, além de ser apoiado numa quantidade de elementos textuais presentes nas sentenças e no contexto das mensagens (SAKAKI; OKAZAKI; MATSUO, 2010). Aliás, para o pré-processamento das mensagens os autores aplicaram técnicas de limpeza de dados (por exemplo, filtragem de palavras, remoção das *stop words*, lematização e transformação de dados) (SAKAKI; OKAZAKI; MATSUO, 2010). O classificador treinado pelos autores obteve como resultado do processo de avaliação a precisão de 87,50% na classificação de mensagens relativas a terremotos e tufões, além disso, este mecanismo computacional foi executado em 49.314 *tweets*, onde 6.291 dessas mensagens são relacionadas aos desastres e foram publicadas por 4.218 usuários diferentes (SAKAKI; OKAZAKI; MATSUO, 2010).

Posteriormente, os pesquisadores focaram na aplicação de Filtros de Kalman e Filtros de Partículas para a estimativa da localização de possíveis centros de terremotos e da trajetória de prováveis tufões a partir de tweets (SAKAKI; OKAZAKI; MATSUO, 2010). Assim, os autores com esse mecanismo de identificação de fenômenos naturais obtiveram a acurácia de 89,7% no reconhecimento de terremotos em território japonês com a escala 2 da *Japan Meteorological Agency* (JMA) e 96% de acurácia na identificação de abalos sísmicos classificados como escala 3 da JMA (SAKAKI; OKAZAKI; MATSUO, 2010).

Por último, o *software* desenvolvido pelos autores notifica via endereço eletrônico a população afetada 1 minuto antes da chegada dos fenômenos na região, sendo um resultado admirável, visto que a JMA avisa aos habitantes somente 6 minutos após o desastre natural ter ocorrido via televisão (SAKAKI; OKAZAKI; MATSUO, 2010).

- ***Using social media to enhance emergency situation awareness*** (YIN *et al.*, 2012)

  Nesse trabalho os autores desenvolveram uma abordagem computacional que usufrui de técnicas de PLN e MD com o objetivo de obter a SAW de emergências a partir de mensagens publicadas no *Twitter* e consequentemente fornecer aos profissionais da DC informações para auxiliar a GD (YIN *et al.*, 2012). A abordagem proposta nessa pesquisa possui diversos módulos, como: extração de dados, detecção de incidentes, classificação textual, agrupamento de informações, *geotagging* e visualização de dados (YIN *et al.*, 2012).

  Primeiramente, o módulo de captação de dados coletou 66 milhões de *tweets* geo localizados de 2,51 milhões usuários do período de 2010 até 2011 de algumas regiões do globo terrestre (por exemplo, Nova Zelândia, Inglaterra, entre outras) via API de *Streamming* do *Twitter*, inclusive para a realização dessa coleta foram executadas diversas pesquisas nos *tweets* com o intuito de identificar a ocorrência de determinadas palavras-chave correlatas a emergências e desastres naturais (YIN *et al.*, 2012). Já o módulo de detecção de incidentes, é responsável pelo monitoramento das informações coletadas do Twitter em tempo real e a

notificação da existência de possíveis emergências no mundo real, aliás esses incidentes são identificados pelo *software* devido à quantidade exacerbada de mensagens relacionadas publicadas na rede social (YIN *et al.*, 2012).

Além disso, o módulo de classificação textual corresponde a um mecanismo computacional capaz de identificar mensagens publicadas no *Twitter* relacionadas com desastres naturais (YIN *et al.*, 2012). Assim, os autores utilizaram para o treinamento do classificador um espaço amostral de 450 *tweets* publicados em fevereiro de 2011 durante o Terremoto de *Christchurch* na Nova Zelândia, no qual foram rotulados manualmente de forma binária quanto ao impacto do fenômeno na infraestrutura da cidade (YIN *et al.*, 2012). Inclusive, as fases do pré-processamento executadas pelos pesquisadores nas mensagens correspondem a filtragem textual, remoção de *stop words*, tokenização e transformação de dados simbólicos para numéricos, além de que essas informações pré-processadas foram utilizadas para o treinamento dos algoritmos de ML chamados NB e SVM, onde alcançaram os resultados de 86,2% e 87,5% de precisão de classificação, respectivamente (YIN *et al.*, 2012).

Ademais, o módulo de agrupamento de informações corresponde a implementação de um algoritmo de clusterização *online*, onde possui a tarefa de agrupar as mensagens publicadas no *Twitter* conforme o fenômeno que elas são relacionadas (YIN *et al.*, 2012). Aliás, para esta etapa foram utilizados 3500 *tweets* publicados em fevereiro de 2011 durante o Terremoto de *Christchurch*, inclusive as mensagens que compõem o espaço amostral dessa etapa foram transformadas para caracteres numéricos via *Term Frequency - Inverse Document Frequency* e agrupadas pelo algoritmo proprietário chamado *Online Incremental Clustering*, resultando num coeficiente de *Silhouette* de 0,42 (YIN *et al.*, 2012).

Por último, o módulo de *geotagging* é responsável pela exibição das informações inerentes as mensagens publicadas no *Twitter* em um mapa, como localização geográfica (latitude e longitude) (YIN *et al.*, 2012). Já o módulo de visualização de dados, disponibiliza diversas interfaces computacionais que possuem o objetivo de propiciar aos usuários do programa de computador maneiras de explorar as informações geradas pelos outros módulos (por exemplo, *tweets* relacionados com emergências, agrupamentos textuais, notificações de incidentes, entre outras) (YIN *et al.*, 2012).

- ***Multimedia Data Fusion for Event Detection in Twitter by Using Dempster-Shafer Evidence Theory*** (ALQHTANI; LUO; REGAN, 2015)

  Nesse trabalho é apresentado uma abordagem de Fusão Multimodal de dados textuais e imagens utilizando a Teoria de Dempster-Shafer, na Língua Inglesa é chamada de *Dempster-Shafer Theory* (DST), com o intuito de detectar eventos, como: terremotos, furacões, alagamentos, entre outros (ALQHTANI; LUO; REGAN, 2015). A DST é uma teoria amplamente aplicada em problemas de classificação de dados, onde possui o objetivo de

representar a incerteza presente nessa categoria de problemas, além de ser uma alternativa para a teoria probabilística tradicional (ALQHTANI; LUO; REGAN, 2015). Inclusive, a aplicação dessa teoria é vantajosa em problemas que há diversas categorias de dados, visto que todos os elementos presentes nas amostras das diferentes mídias são considerados por esta estratégia (ALQHTANI; LUO; REGAN, 2015).

O espaço amostral utilizado pelos autores para o desenvolvimento dessa pesquisa corresponde as mensagens publicadas no *Twitter* de 25 de agosto de 2014 até 30 de agosto de 2014 na cidade de Napa no estado da Califórnia nos Estados Unidos da América (EUA) (ALQHTANI; LUO; REGAN, 2015). Dessa forma, para a execução da etapa de MT dos *tweets* os pesquisadores aplicaram os processos de limpeza de texto e transformação de dados, onde na primeira etapa foram filtradas as mensagens escritas na Língua Inglesa, houve a conversão de todos os elementos das sentenças para minúsculo e foram removidas as *stop words* das mensagens, já na segunda fase, foi aplicado o método de transformação de informações chamado TF-IDF (ALQHTANI; LUO; REGAN, 2015).

Além disso, para o processamento das imagens contidas nas mensagens publicadas no *Twitter*, foram obtidos os pontos-chave das fotografias automaticamente via *Scale-invariant feature transform* (SIFT), posteriormente esses recursos visuais são agrupados com o auxílio do algoritmo K-Means, resultando num dicionário com o intuito de descrever os diversos padrões de imagens (ALQHTANI; LUO; REGAN, 2015). Assim, com o mapeamento dos pontos-chave para palavras visuais é possível retratar uma imagem como um "saco de palavras visuais" e consequentemente converter esse saco de palavras em um vetor de palavras visuais (ALQHTANI; LUO; REGAN, 2015).

Por último, os pesquisadores realizaram experimentos com o intuito de avaliar quão precisa é a combinação de informações textuais e visuais para a detecção de eventos, portanto após o processamento dos dados treina-se o algoritmo de DST para as informações isoladas e combinadas ao nível de recurso (Fusão Prévia) (ALQHTANI; LUO; REGAN, 2015). Desse modo, o modelo treinado com os dados textuais atingiram 93% de precisão, já o modelo com as informações imagéticas alcançaram 86% de precisão e o modelo treinado com as informações textuais e imagéticas combinadas atingiu 97% de precisão (ALQHTANI; LUO; REGAN, 2015).

- ***Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery*** (HUANG; XIAO, 2015)

  De acordo com Huang e Xiao (2015), nesse estudo apresenta-se uma abordagem de MD que possui o objetivo de identificar automaticamente as mensagens do *Twitter* relacionadas com desastres naturais e auxiliar a GD nas etapas de preparação, resposta e recuperação. Aliás, segundo Huang e Xiao (2015) esta abordagem não auxilia a fase de mitigação da GD, pois a etapa de mitigação é focada na prevenção ou minimização dos impactos dos fenômenos naturais a longo prazo. Inclusive, o estudo de caso dessa pesquisa é o

furacão Sandy que ocorreu em outubro de 2012 nos EUA, desse modo os *tweets* geo localizados utilizados nesse artigo pertencem à cidade de Nova York (NY) nos EUA e foram publicados no intervalo de 10 de outubro de 2012 até 27 de novembro de 2012, onde corresponde ao período de acontecimento desse fenômeno natural (HUANG; XIAO, 2015).

Ademais, os autores desse trabalho coletaram cerca de 1,763 milhão de *tweets* disponíveis na plataforma Gnip[1] e pré-processaram esses dados, dessa forma na fase de limpeza textual da etapa de pré-processamento os pesquisadores executaram as seguintes estratégias: filtragem textual por palavras-chave relacionadas com desastres naturais e fases da GD, remoção das *stop words* e tokenização (HUANG; XIAO, 2015). Já na fase de transformação de dados, as informações textuais foram transformadas de caracteres simbólicos para numéricos (HUANG; XIAO, 2015). Logo após esses procedimentos, os autores obtiveram cerca de 8.807 mensagens publicadas no *Twitter* pré-processadas, onde foram submetidas a um processo de classificação manual quanto ao relacionamento com as etapas da GD (HUANG; XIAO, 2015).

Posteriormente aos dados serem rotulados, os pesquisadores empregaram essas informações para treinar e testar os seguintes algoritmos de ML: *K-Nearest Neighbors* (KNN), LR e NB (HUANG; XIAO, 2015). Inclusive, as seguintes métricas de avaliação foram utilizadas pelos autores para analisar o desempenho dos algoritmos: precisão, recall e F1-Score (HUANG; XIAO, 2015). Dessa forma, a melhor solução encontrada foi o classificador baseado em LR com precisão de 64,7%, *recall* de 71,1% e *f1-score* 66,4%.

Por último, os autores realizaram uma análise espaço-temporal com as informações obtidas do mecanismo de classificação para identificar o impacto que os *tweets* exercem na SAW de desastres naturais (HUANG; XIAO, 2015). Dessa forma, os pesquisadores fizeram algumas constatações sobre os resultados das análises, sendo a primeira relacionada ao acréscimo de *tweets* relativos à etapa de preparação da GD nas redes sociais antes de a região de NY ser acometida pelo furacão Sandy, inclusive muitos cidadãos foram motivados a expressar suas opiniões e preocupações no *Twitter* devido às declarações de emergência emitidas pelo presidente dos EUA daquele período (Barack Obama) (HUANG; XIAO, 2015). Já a segunda constatação, relaciona-se com o conteúdo das mensagens pertencentes a categoria recuperação, visto que foi observado pelos autores que grande parte dos *tweets* destas classes foram publicados em parques públicos (por exemplo, *Central Park*) e exibem diversas fotografias de árvores caídas, assim nota-se a eficácia na obtenção de SAW de desastres naturais a partir de *tweets* pelo mecanismo de classificação proposto (HUANG; XIAO, 2015).

- ***Mining Multimodal Information on Social Media for Increased Situational Awareness*** (KELLY; ZHANG; AHMAD, 2017)

[1] url: <http://gnip.com/>

Nesse trabalho os autores desenvolveram uma abordagem de combinação de dados textuais e imagéticos com o intuito de auxiliar a etapa de resposta da GD, além disso, os autores compararam as informações combinadas com dados meteorológicos e disponibilizaram uma ferramenta *online* capaz de proporcionar a visualização das informações analisadas (KELLY; ZHANG; AHMAD, 2017).

Para a criação do mecanismo computacional capaz de classificar as mensagens advindas do *Twitter*, primeiramente os autores extraíram os 7.944 *tweets* de 20 dezembro de 2015 até 2 janeiro de 2016 da região da Irlanda via API pública do *Twitter*, visto que neste período nesta região há diversas ocorrências de tempestades e inundações (KELLY; ZHANG; AHMAD, 2017). Desse modo, para o processamento textual os pesquisadores empregaram a biblioteca *CoreNLP* desenvolvida na universidade de *Stanford* para a realização da filtragem e a tokenização das sentenças escritas na Língua Inglesa, inclusive a filtragem textual realizada pelos autores foi baseada em um dicionário de domínio de desastres naturais, resultando em 484 *tweets* relacionados aos fenômenos naturais encontrados (KELLY; ZHANG; AHMAD, 2017). Já para o processamento das imagens contidas nos *tweets*, foi empregado um modelo de reconhecimento de imagens treinado com o intuito de identificar categorias de fotos relacionadas com desastres naturais e descrevê-las (KELLY; ZHANG; AHMAD, 2017). Assim, segundo Kelly, Zhang e Ahmad (2017) as informações textuais pré-processadas (nível de recurso) e os dados imagéticos descritos pelo mecanismo de classificação de fotografias (nível de decisão) são combinados e treina-se um SVM com as informações combinadas (Fusão Multimodal de modo Híbrido) com o intuito de obter a SAW de desastres naturais a partir dos *tweets*.

Por último, os pesquisadores afirmam que a combinação do conteúdo textual e imagético das mensagens advindas do *Twitter* melhora o desempenho do mecanismo de classificação de *tweets* quanto a identificação de mensagens relacionadas a fenômenos naturais (KELLY; ZHANG; AHMAD, 2017). Ademais, os autores demonstram que em experimentos realizados houve o acréscimo do volume de dados relacionados a tempestades e inundações publicados no *Twitter* em períodos que houve o acontecimento desses fenômenos na Irlanda, inclusive a correlação entre as séries temporais de mensagens do *Twitter* e a de dados meteorológicos é de 54%, então evidencia-se que o sistema desenvolvido pelos autores pode detectar conversas sobre fenômenos naturais do mundo real no ambiente virtual (KELLY; ZHANG; AHMAD, 2017).

- ***Extraction of pluvial flood relevant volunteered geographic information(VGI) by deep learning from user generated texts and photos*** (FENG; SESTER, 2018)

  De acordo com Feng e Sester (2018), nesse trabalho apresenta-se uma abordagem de identificação de VGI's relacionadas com inundações a partir de textos e imagens utilizando técnicas de PLN e DL. Inclusive, essa pesquisa possui o objetivo de obter SAW de desastres naturais a partir de mensagens de redes sociais, além de consequentemente orientar aos

profissionais responsáveis pela GD a localização geográfica das pessoas afetadas pelos fenômenos naturais (FENG; SESTER, 2018).

Primeiramente, para o processamento das informações textuais os autores desenvolveram um mecanismo computacional capaz de identificar mensagens do *Twitter* relacionadas a alagamentos, assim 3,6 milhões de *tweets* geo localizados publicados no território oeste do continente europeu, escritos em diversos idiomas nativos daquela região (por exemplo, a Língua Inglesa, Francesa, Germânica, Espanhola, entre outras) e contidos no período de 1 de junho de 2016 até 30 de junho de 2016 foram captados via API de *Streamming* do *Twitter* (FENG; SESTER, 2018).

Posteriormente, os *tweets* capturados pelos pesquisadores foram submetidos a etapa de pré-processamento da MT, onde a princípio na fase de limpeza de dados cujo objetivo é remover ruídos, foram aplicadas as seguintes técnicas: filtragem textual, ou seja, a busca por mensagens contendo as palavras-chave relacionadas a alagamentos; remoção de endereços eletrônicos, pontuações, *stop words* e mensagens idênticas; *stemming*, ou seja, redução das palavras ao radical (por exemplo, inundação o radical é inundar) (FENG; SESTER, 2018). Já na fase de transformação de dados da etapa de pré-processamento cujo objetivo é transformar os dados compreensíveis para as máquinas, os autores aplicaram as técnicas de TF-IDF e *Word Embeddings* com o intuito de converter os caracteres simbólicos das mensagens para numéricos (FENG; SESTER, 2018).

Logo após as informações textuais serem pré-processadas, os pesquisadores rotularam automaticamente as mensagens do *Twitter* de forma binária quanto ao relacionamento do conteúdo com inundações (por exemplo, positivo (1), caso as mensagens sejam relacionadas com alagamentos e negativo (0), caso contrário) (FENG; SESTER, 2018). Assim, segundo Feng e Sester (2018) a base de treinamento gerada para o mecanismo de classificação textual corresponde a 79 938 *tweets*, sendo 50% das observações positivas e 50% negativas.

Portanto, para o treinamento do classificador de *tweets*, os pesquisadores aplicaram a técnica de *Grid Search* presente na biblioteca chamada Scikit-Learn com o objetivo de otimizar os parâmetros dos algoritmos RF, LR, SVM e CNN (FENG; SESTER, 2018). Assim, após a obtenção dos melhores parâmetros, os autores executaram os processos de validação cruzada nos algoritmos com o intuito de identificar a melhor solução de classificação textual para o espaço amostral analisado (FENG; SESTER, 2018). Desse modo, segundo (FENG; SESTER, 2018) a melhor solução encontrada foi a que utiliza CNN alcançando uma precisão de 78,68% na classificação dos dados, inclusive esta abordagem utiliza *Word Embeddings* do tipo Word2Vec como estratégia de transformação de dados.

Aliás, para o desenvolvimento do mecanismo de classificação de imagens desse trabalho, os autores coletaram 22 800 imagens do período de 1 de julho de 2016 até 28 de outubro

de 2016 de três diferentes categorias, sendo elas fotografias com e sem relacionamento com alagamentos e fotografias relacionadas com superfícies aquáticas (FENG; SESTER, 2018). Dessa forma, após a captação das imagens, os pesquisadores aplicaram a estratégia de Aprendizado por Transferência, do inglês *Transfer Learning* (TL), no espaço amostral imagético com o objetivo de se beneficiarem da experiência de treinamento de uma CNN consolidada na literatura, inclusive a CNN utilizada neste experimento para apoiar o processo de identificação de imagens é a GoogLeNet (modelo Inception-V3) presente na biblioteca do *Tensorflow* e treinada com 1,2 milhões de fotografias de 100 diferentes categorias (FENG; SESTER, 2018).

Posteriormente, segundo Feng e Sester (2018) os pesquisadores removeram a última camada da GoogleLeNet e obtiveram como saída diversos vetores de característica com 2048 valores pertencentes a cada imagem. Portanto, após esse procedimento, os autores dividiram a base de dados de imagens em 90% para treino e 10% para validação, treinaram e testaram os algoritmos de ML RF, LR, MLP, xgboost e *Gradient Boosted Tree* (GBT) com os resultados das informações processados pela GoogleLeNet e obtiveram como melhor classificador de imagens o xgboost com 92,95% de precisão (FENG; SESTER, 2018).

Além disso, de acordo com Feng e Sester (2018) neste trabalho foi aplicado algoritmo de agrupamento *Spatio-Temporal Density-Based Spatial Clustering of Applications with Noise* (ST-DBSCAN) com o intuito de identificar a região de ocorrência de fenômenos naturais, inclusive a distância de formação de grupos que esse algoritmo utilizou (8 km) foi obtida empiricamente após a análise de diversos resultados derivados do tamanho de célula de chuva intensa do Reino Unido (10 km). Dessa forma, os autores alcançaram o resultado de 99% de confiança da existência de fenômenos naturais em áreas específicas a partir de *tweets* relacionados com inundações (FENG; SESTER, 2018). Ademais, os pesquisadores desenvolveram uma plataforma que exibe os *tweets* geo localizados relacionados a alagamentos num mapa iterativo e a região com a possibilidade da ocorrência de desastres naturais conforme as informações analisadas das redes sociais (FENG; SESTER, 2018).

Por último, como os mecanismos de classificação de texto e imagem foram treinados separadamente, então são realizados dois experimentos para validar a classificação das informações de modo independente e combinado (Fusão Tardia), inclusive as informações utilizadas nesta etapa foram captadas de 17 de maio de 2016 até 10 de junho de 2016 via API de *Streamming* do *Twitter* (FENG; SESTER, 2018). Sendo que, no primeiro experimento foram utilizadas as observações de Paris, logo os pesquisadores encontraram a maior correlação com os dados meteorológicos nas informações textuais (0,492). Já no segundo experimento, foram empregados dados de Londres, assim os autores descobriram a maior correlação com as informações climáticas nas imagens (0,836) (FENG; SESTER, 2018). Dado o exposto, segundo Feng e Sester (2018) conclui-se que existe

uma correlação considerável entre as observações meteorológicas e os *tweets*, além de que a abordagem desenvolvida pode identificar inundações pluviais, no entanto, a proposta apresenta dificuldades para identificar inundações fluviais (advindas de rios).

- ***Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network*** (BRUIJN *et al.*, 2020).

  Nessa pesquisa é apresentado o desenvolvimento de uma *Convolutional Neural Network* (CNN), na Língua Portuguesa é chamada de Rede Neural Convolucional, que possui o objetivo de melhorar o processo de classificação textual para a detecção de alagamentos a partir da Fusão Multimodal de informações textuais e meteorológicas (BRUIJN *et al.*, 2020).

  Primeiramente, para a criação do espaço amostral textual desse trabalho, foi realizada a coleta de 56,8 milhões de *tweets* do período de 29 de julho de 2014 até 11 de novembro de 2018 via API de *Streamming* do *Twitter*, inclusive essas mensagens foram submetidas a um processo de filtragem de texto por palavras-chave relacionadas a inundações em quatro idiomas (Inglês, Francês, Espanhol e Indonésio) (BRUIJN *et al.*, 2020). Aliás, para a identificação da geolocalização dos *tweets*, os autores aplicaram no espaço amostral textual o algoritmo *Toponym-based Algorithm for Grouped Geoparsing of Social media* (TAGGS), no qual possui o objetivo de identificar a latitude e a longitude das mensagens em todo globo terrestre com base em dados de referência (por exemplo, nomes de ruas, códigos postais, entre outros) (BRUIJN *et al.*, 2020). Assim, o algoritmo TAGGS identificou as coordenadas geográficas de 6,84 milhões *tweets*, no qual foram alocados a diversas sub-bacias hidrográficas obtidas do conjunto de dados HydroBASINS (BRUIJN *et al.*, 2020).

  Posteriormente, para a classificação manual das mensagens contidas no espaço amostral textual, os autores excluíram os conteúdos idênticos, indisponíveis e que não detalham os locais geográficos, além de rotularem manualmente de modo binário 1000 *tweets* de cada idioma em “relevantes”, caso relacionados a inundações e “irrelevantes”, caso contrário (BRUIJN *et al.*, 2020). Além disso, segundo Bruijn *et al.* (2020) para o pré-processamento das informações textuais foram aplicadas as etapas de limpeza de dados e transformação de dados, onde na primeira fase foram removidos os endereços eletrônicos contidos nas mensagens, aplicado o processo de filtragem de texto, realizado a remoção de *stop words* e executado o processo de tokenização, já na segunda fase os elementos das sentenças pré-processadas são transformados de caracteres simbólicos para caracteres numéricos via *Word Embeddings* do tipo *Fast Text*.

  Para a criação do espaço amostral meteorológico, os autores captaram as informações climáticas do conjunto de dados HydroBASINS do período de 2009 até 2018, assim foram coletados os dados históricos de precipitação das sub-bacias hidrográficas conforme as

datas e os horários de publicação dos *tweets* e a quantidade de chuva acumulada nas regiões das sub-bacias (BRUIJN *et al.*, 2020).

Os autores dessa pesquisa para o desenvolvimento da CNN, a princípio projetaram uma sub-rede textual que usufrui de diversas características contidas no espaço amostral para identificar os tópicos das mensagens do *Twitter*, inclusive esta sub-rede contém uma série de nós (não binários) (BRUIJN *et al.*, 2020). Ademais, os pesquisadores também desenvolveram uma sub-rede hidrológica que utiliza as diversas informações meteorológicas contidas no espaço amostral para identificar a existência de inundações nas observações, aliás essa sub-rede contém somente um nó (não binário) (BRUIJN *et al.*, 2020). Logo após, foi desenvolvida uma sub-rede de decisão responsável pela combinação das informações produzidas pela sub-rede textual e sub-rede hidrológica (Fusão Multimodal de modo Híbrido) com o intuito de classificar as mensagens em relacionadas a alagamentos ou não (BRUIJN *et al.*, 2020).

Por último, os autores compararam o desempenho da CNN quanto a inclusão de informações meteorológicas, dessa forma alcançaram 91% precisão (BRUIJN *et al.*, 2020). Além disso, os pesquisadores desse trabalho afirmam que a utilização de informações climáticas melhora a capacidade de aprendizado da CNN até para mensagens escritas em idiomas não contidos no espaço amostral, logo minimiza a necessidade de as mensagens do *Twitter* serem classificadas manualmente para um novo idioma (BRUIJN *et al.*, 2020).

3.2 Síntese e Comparação dos Trabalhos Relacionados

A Tabela 3 apresenta uma síntese dos trabalhos correlatos detalhados na Seção 3.1. Desse modo, observa-se nesta tabela a descrição e as categorias das informações utilizadas por cada trabalho. Além disso, na Tabela 3 há também a apresentação dos tipos de estratégia de Fusão Multimodal que foram empregadas, caso as pesquisas usufruam de abordagens de combinação de informações midiáticas. Aliás, nesta mesma tabela, são exibidas as abordagens utilizadas pelos trabalhos correlatos para a identificação de áreas suscetíveis a ocorrência de fenômenos naturais.

Os trabalhos correlatos apresentados na Seção 3.1 e sintetizados na Tabela 3 trazem abordagens computacionais para a obtenção de Consciência Situacional de desastres naturais a partir de *tweets* com o intuito de auxiliar as diversas etapas da GD. Desse modo, a partir desse levantamento bibliográfico é possível observar que na maioria dos trabalhos os pesquisadores utilizaram a estratégia de classificação manual dos *tweets* para a criação dos rótulos, inclusive as mensagens foram categorizadas quanto a presença de relacionamento com desastres naturais ou etapas da GD (por exemplo, Yin *et al.* (2012), Huang e Xiao (2015), Feng e Sester (2018), Bruijn *et al.* (2020), entre outros). Assim, essa também é a estratégia utilizada nesta pesquisa para a criação das classes do espaço amostral, visto que para o desenvolvimento do modelo de

Tabela 3 – Síntese dos trabalhos correlatos encontrados na literatura

Referência	Tipos de Dados	Algoritmos de ML	Tipo de Fusão Multi-modal	Identificação das áreas de ocorrência de fenômenos naturais	Auxílio em quais etapas da GD
Sakaki et al. (2010)	Tweets	SVM	Não	Filtro de Kalman e Filtro de Partículas	Preparação e Resposta
Yin et al. (2012)	Tweets	NB, SVM Online Incremental Clustering	Não	Não	Resposta
Alqhtani et al. (2015)	Tweets e Imagens	K-Means	Prévia	Não	Resposta
Huang e Xiao (2015)	Tweets	KNN, NB e LR	Não	Não	Preparação, Resposta e Recuperação
Kelly et al. (2017)	Tweets, Imagens e Informações Meteorológicas	SVM	Híbrida	Não	Resposta
Feng e Sester (2018)	Tweets, Imagens e Informações Meteorológicas	SVM, NB, DT, LR, CNN, MLP, GBT, xgboost, ST-DBSCAN	Tardia	Agrupamento espaço-temporal de tweets	Resposta
Bruijn et al. (2020)	Tweets e Dados Meteorológicos	CNN	Híbrida	Não	Resposta
Esta dissertação	**Tweets, Informações Meteorológicas e Ocorrências Históricas de Alagamentos**	**RF, DT, SVM, NB, LR, DBSCAN, OPTICS, Agglomerative Clustering**	**Prévia, Tardia, Híbrida**	**Agrupamento de ocorrências históricas de alagamentos**	**Resposta**

Fonte: Elaborada pelo autor.

Fusão Multimodal capaz de identificar possíveis vítimas de alagamentos há a necessidade da identificação prévia dos *tweets* relacionados com situações climáticas.

Além disso, observa-se nos trabalhos relacionados que a maioria das abordagens computacionais propostas utilizaram as estratégias de limpeza de dados e transformação de dados presentes no processo de pré-processamento da metodologia de MT. Sendo que, nota-se que para a execução da limpeza de dados os autores frequentemente empregaram as técnicas de filtragem de dados, remoção de *stop words* e tokenização (por exemplo, Sakaki, Okazaki e Matsuo (2010),

Alqhtani, Luo e Regan (2015), Huang e Xiao (2015), Feng e Sester (2018), Bruijn *et al.* (2020), entre outros). Ademais, para a aplicação da transformação de dados alguns trabalhos empregaram a técnica de TF-IDF (por exemplo, Yin *et al.* (2012), Alqhtani, Luo e Regan (2015), Feng e Sester (2018)), já outras pesquisas utilizaram técnicas baseadas em DL, como *Word Embeddings* (por exemplo, Feng e Sester (2018), Bruijn *et al.* (2020)). Dessa forma, nesta dissertação para a etapa de limpeza de dados, utiliza-se as mesmas técnicas dos trabalhos correlatos e há a inclusão de um mecanismo de correção de palavras escritas coloquialmente. Já na fase de transformação de dados, existe a utilização das seguintes técnicas: BOW, TF-IDF e *Word Embenddings*, além da avaliação do impacto no desempenho dos algoritmos de ML quanto a execução desses métodos no espaço amostral.

Aliás, percebe-se nos trabalhos correlatos que para a classificação das informações do espaço amostral alguns algoritmos foram escolhidos predominantemente, como: SVM, NB, DT e LR. Desse modo, nesta dissertação para a mesma função escolhemos os algoritmos SVM, NB, DT, RF e LR, onde segundo Faceli *et al.* (2011) somente o DT e o RF compartilham o mesmo paradigma, já os demais são distintos. Além disso, nota-se que para o agrupamento das informações geográficas a pesquisa de Feng e Sester (2018) utilizou o ST-DBSCAN. Dessa maneira, nesta dissertação foram selecionados os seguintes algoritmos para a função de agrupamento de dados geográficos: DBSCAN, OPTICS, *Agglomerative Clustering*. Inclusive, de acordo com Ankerst *et al.* (1999), Borah e Bhattacharyya (2004) o DBSCAN e o OPTICS são algoritmos que compartilham o mesmo paradigma de agrupamento, ou seja, o baseado em densidade, já o *Agglomerative Clustering* pertence ao paradigma de agrupamento baseado em hierarquia (FACELI *et al.*, 2011).

Ademais, nota-se que uma grande parcela das pesquisas correlatas encontradas na literatura são focadas na obtenção de SAW de mensagens publicadas no *Twitter* para o auxílio na etapa de resposta da GD, visto que esta fase é acionada durante e depois o acontecimento dos desastres naturais e concentra-se nos processos de busca e salvamento da população afetada (POSER; DRANSCH, 2010). Além disso, percebe-se que os trabalhos correlatos descobertos na literatura obtêm SAW de nível 2, ou seja, a partir das mensagens do *Twitter* é possível compreender o estado atual dos desastres naturais e auxiliar a tomada de decisão dos profissionais responsáveis pela GD. Dessa maneira, a abordagem desenvolvida nesta dissertação é focada em auxiliar a etapa de resposta da GD e propiciar a obtenção de SAW de nível 2.

Os trabalhos de Sakaki, Okazaki e Matsuo (2010), Yin *et al.* (2012), Alqhtani, Luo e Regan (2015), Huang e Xiao (2015) utilizam somente as informações advindas dos *Twitter* para a construção dos modelos de classificação de mensagens relacionadas a desastres naturais. Desse modo, para que seja possível a construção de modelos mais precisos do que os presentes nesses trabalhos é necessário a combinação dos *tweets* com informações externas (por exemplo, imagens, dados de sensores, entre outros) (YIN *et al.*, 2012; HUANG; XIAO, 2015). Portanto, nesta dissertação há a combinação de informações textuais, meteorológicas e geográficas para a produ-

ção de uma abordagem computacional capaz de obter SAW de alagamentos mais precisamente que as demais versões unimodais presentes na literatura.

Os trabalhos de Kelly, Zhang e Ahmad (2017), Feng e Sester (2018) usufruem das seguintes categorias de dados: textuais, imagéticos e climáticos. Assim, os dois primeiros dados foram empregados para a construção dos modelos de identificação de informações relacionadas com alagamentos, já o terceiro foi utilizado somente para validar as soluções elaboradas pelos pesquisadores. Dessa forma, segundo Feng e Sester (2018) essas abordagens desenvolvidas estão sujeitas a utilização de informações incertas, pois todas as características empregadas pelos modelos são VGI, inclusive VGI falsificados são constantemente publicados pelas pessoas nas redes sociais de forma intencional ou involuntária. Portanto, a combinação dos dados textuais e imagéticos ao nível de recurso ou decisão com os dados meteorológicos torna os modelos mais precisos (HUANG; XIAO, 2015). Inclusive, nesta dissertação investiga-se a combinação de dados de diferentes mídias com informações contextuais para o desenvolvimento de um modelo de obtenção de SAW de inundações mais preciso.

Além disso, no trabalho de Feng e Sester (2018) explora-se o agrupamento de *tweets* de forma espaço-temporal via ST-DBSCAN com o objetivo de reconhecer as regiões propícias para o acontecimento de desastres naturais conforme o fluxo de mensagens publicadas no *Twitter*. Inclusive, segundo Feng e Sester (2018) a metodologia escolhida nesse trabalho para a definição da melhor distância de formação de grupos foi baseada na combinação e verificação de diversos agrupamentos visualmente, no entanto, essa estratégia é passível a erros ocasionados pelos seres-humanos, visto que não há um consenso de diversos juízes sobre os agrupamentos gerados e não são aplicadas métricas de avaliação de formação de grupos. Desse modo, nesta dissertação exploram-se as estratégias empíricas e estatísticas de definição de distância de formação de agrupamentos, onde a melhor solução é definida de acordo com o valor resultante da métrica de avaliação de agrupamentos chamada *Silhouette*.

Ademais, no trabalho de Bruijn *et al.* (2020) é elaborado uma abordagem baseada em CNN para a obtenção de SAW de desastres naturais a partir de dados textuais e contextuais. Desse modo, para a execução do processo de fusão espaço-temporal das informações meteorológicas com os *tweets*, os autores empregaram um algoritmo chamado TAGGS para descobrir a localização das mensagens de forma semântica, no entanto, essa estratégia gera algumas geo localizações incertas que comprometem a eficácia do modelo (BRUIJN *et al.*, 2020). Portanto, nesta dissertação são utilizados somente *tweets* geo localizados, assim evita-se a presença de localizações geográficas incertas e consequentemente de combinações espaço-temporal entre *tweets* e dados climáticos errôneas.

Por último, no trabalho de Bruijn *et al.* (2020) as seguintes informações contextuais foram utilizadas: precipitação e quantidade de chuva acumulada. Assim, de acordo com os autores desse artigo, o processo de adicionar outros dados contextuais (por exemplo, umidade, altura das tempestades, entre outros) ao modelo torna o mecanismo computacional mais preciso

(BRUIJN *et al.*, 2020). Desse modo, nesta dissertação utilizam-se como variáveis contextuais as seguintes informações meteorológicas: precipitação, umidade, temperatura, temperatura do ponto do orvalho e pressão atmosférica, pois o objetivo deste trabalho é a produção de um mecanismo computacional que capta SAW de inundações de maneira precisa a partir de *tweets* e dados contextuais.

3.3 Considerações Finais

Este capítulo apresentou alguns trabalhos relacionados encontrados na literatura que usufruem de informações advindas de redes sociais para a obtenção de SAW de alagamentos e auxílio a GD. Inclusive, foi apresentada uma síntese dos trabalhos correlatos e uma comparação entre as pesquisas relacionadas com a abordagem de combinação de dados heterogêneos desenvolvida nesta dissertação. Aliás, foi constatado que existem trabalhos correlatos que utilizam para o desenvolvimento dos modelos de Fusão Multimodal somente VGI, além de usufruírem de localizações geográficas passíveis a erros e definirem as regiões propícias ao acontecimento de enchentes sem embasamento estatístico. Portanto, nesta dissertação há a utilização de informações meteorológicas além da VGI na abordagem de Fusão Multimodal, além disso, neste trabalho utiliza-se dados geográficos certeiros e embasamento estatístico na descoberta das áreas propícias ao acontecimento de alagamentos.

CAPÍTULO

4

MODELO DE FUSÃO MULTIMODAL

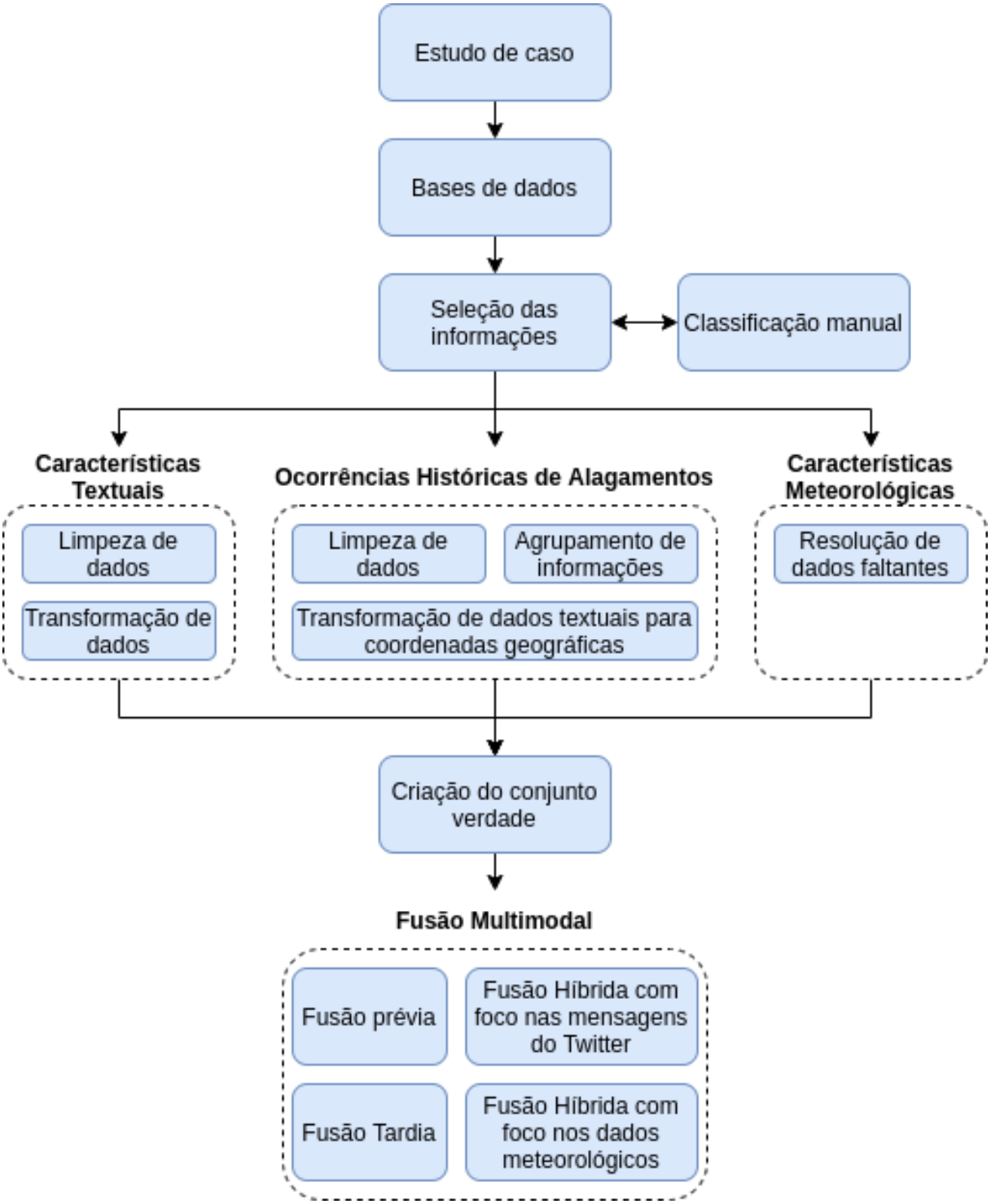
Este capítulo possui o intuito de apresentar uma abordagem de Fusão Multimodal capaz de obter a SAW de desastres naturais e auxiliar a etapa de resposta da GD. Desse modo, na Seção 4.1 é relatado a metodologia empregada para a confecção do modelo, já na Seção 4.2 são expostas às características dos experimentos necessários para a criação dos modelos de Fusão Multimodal, inclusive nesta abordagem há a realização de três experimentos: descoberta das regiões propícias à ocorrência de alagamentos na cidade de São Paulo; treinamento e avaliação dos modelos de Fusão Multimodal; comparação entre os modelos de Fusão Multimodal e averiguação do impacto da inclusão de dados contextuais na abordagem. Por último, na Seção 4.3 são apresentadas as considerações finais deste capítulo.

4.1 Abordagem

Nesta seção, apresenta-se a descrição do processo de elaboração do modelo de Fusão Multimodal capaz de obter SAW de alagamentos a partir de *tweets* e informações contextuais, além de auxiliar a etapa de resposta da GD. Inclusive, esta abordagem de Fusão Multimodal foi elaborada com base em alguns conceitos empregados por determinados trabalhos correlatos encontrados na literatura (por exemplo, Sakaki, Okazaki e Matsuo (2010), Feng e Sester (2018), Bruijn *et al.* (2020)). Dessa forma, na Figura 17 nota-se a representação visual das fases deste processo.

De acordo com a Figura 17, primeiramente, na abordagem elaborada para a criação do modelo de Fusão Multimodal são discutidos os motivos da cidade de São Paulo ser escolhida como objeto de estudo desta dissertação (Subseção 4.1.1). Logo após, são relatados os procedimentos empregados para: captação das mensagens publicadas no *Twitter*; extração das informações meteorológicas da cidade de São Paulo na plataforma do Instituto Nacional de Meteorologia

Figura 17 – Metodologia para a criação do modelo de Fusão Multimodal capaz de obter SAW de inundações na cidade de São Paulo



Fonte: Elaborada pelo autor.

(INMET)[1]; obtenção das ocorrências históricas de alagamentos da cidade de São Paulo no portal do Centro de Emergências Climáticas da Prefeitura de São Paulo (CGE-SP)[2](Subseção 4.1.2).

[1] url: <https://portal.inmet.gov.br/>
[2] url: <https://www.cgesp.org/v3/>

Posteriormente segundo a Figura 17, são relatadas as estratégias para a seleção das informações utilizadas pelo modelo de Fusão Multimodal (Subseção 4.1.3). Assim, as informações textuais resultantes do processo de seleção de dados são submetidas para a etapa de classificação manual, onde os *tweets* são classificados manualmente por três juízes quanto a presença de relacionamento com inundações (por exemplo, atribui-se positivo (1) para mensagens do *Twitter* relacionadas com chuvas ou inundações, caso contrário atribui-se (0)) (Subseção 4.1.3).

Em seguida de acordo com a Figura 17, apresenta-se a etapa de execução do pré-processamento textual e a exibição das técnicas presentes nas fases de limpeza de dados e transformação de dados (Subsubseção 4.1.5.1). Ademais, há a aplicação da etapa de pré-processamento das ocorrências históricas de alagamentos e a apresentação do processo de limpeza dessas informações, aliás nesta etapa existe a exibição do processo de transformação de ocorrências históricas em coordenadas geográficas e detalhamento do procedimento para a criação de áreas de alagamentos da cidade de São Paulo (Subsubseção 4.1.5.2). Além disso, há a efetuação da etapa de pré-processamento das informações meteorológicas, onde se aplicam estratégias para a resolução de dados faltantes (Subsubseção 4.1.5.3).

Logo após segundo a Figura 17, as informações textuais, meteorológicas e ocorrências históricas de alagamentos resultantes das respectivas etapas de pré-processamento são combinadas e há a criação do "conjunto verdade" deste trabalho, no qual é empregado para o treinamento e teste dos modelos de Fusão Multimodal (Subseção 4.1.6). Por último, são exploradas as abordagens de Fusão Multimodal ao nível de recurso (Fusão Prévia) e ao nível de decisão (Fusão Tardia), além de que são elaborados nesta etapa dois novos modelos de Fusão Multimodal Híbridos, onde o foco de um modelo está nos *tweets*, já o outro é focado nos dados meteorológicos (Subseção 4.1.7).

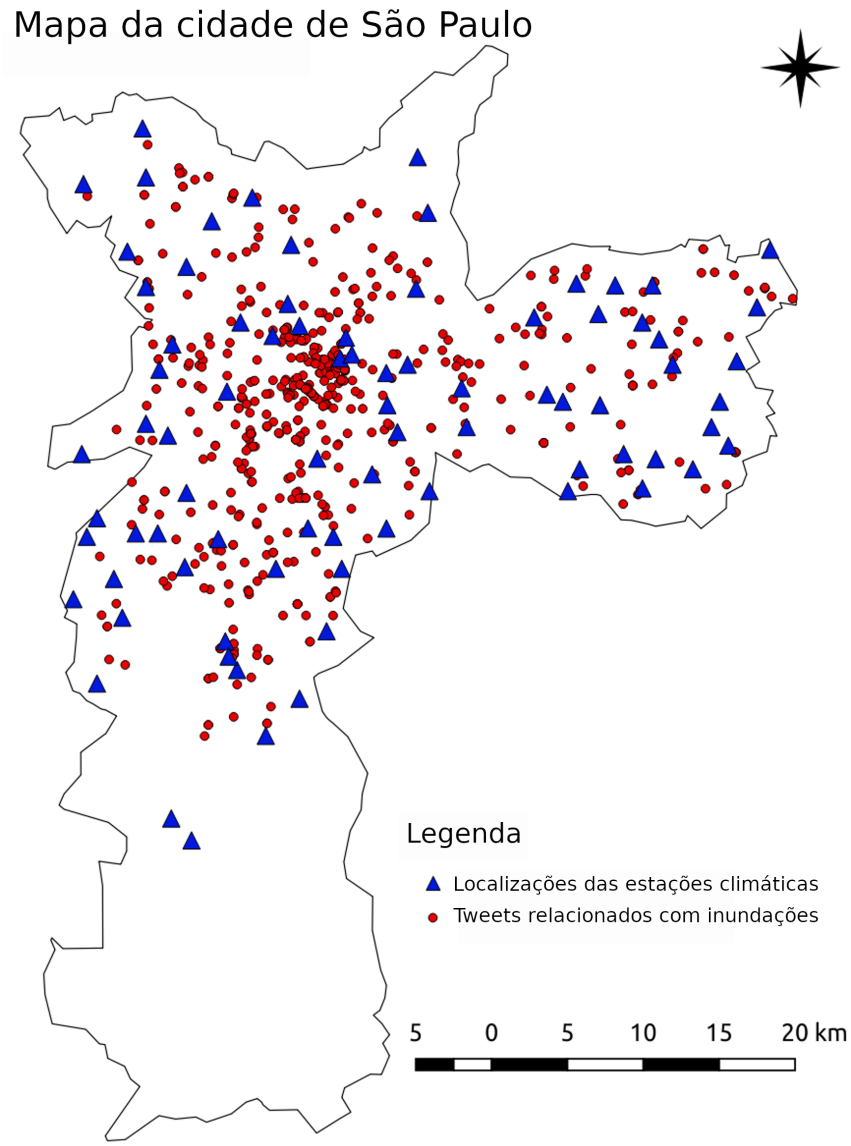
4.1.1 Estudo de Caso

A cidade de São Paulo é o objeto de estudo de caso desta pesquisa, pois se estima que a capital paulista tenha uma população de cerca de 12 milhões de habitantes, onde 674.329 pessoas são residentes de áreas suscetíveis ao acontecimento de inundações (IBGE, 2010). Desse modo, a utilização de abordagens computacionais que auxiliem a etapa de resposta da GD por profissionais da DC, possibilita a minimização dos impactos dos alagamentos na qualidade de vida de uma parcela dos moradores da cidade de São Paulo.

Além disso, de acordo com Andrade *et al.* (2017) o município São Paulo é caracterizado por ser uma das cidades mais populosas da América do Sul, consequentemente é uma das capitais com a maior quantidade de usuários do *Twitter*, logo diariamente há a publicação de uma quantia exacerbada de mensagens na plataforma relacionadas com fenômenos naturais. Ademais, conforme a pesquisa realizada por Andrade *et al.* (2017) com *tweets* relacionados a alagamentos do município de São Paulo no período de janeiro de 2016, os autores constataram que as informações publicadas no *Twitter* tendem a ser agrupar próximas de estações meteorológicas, assim

possibilita-se a aplicação de uma abordagem de combinação espaço-temporal entre os *tweets* e os dados meteorológicos mais assertiva. Portanto, na Figura 18 nota-se esse comportamento. Inclusive, as setas azuis são as diversas localizações das estações meteorológicas da capital paulista, já os pontos vermelhos são os *tweets* relacionados com inundações.

Figura 18 – Mapa da cidade de São Paulo com a localização dos *tweets* relacionados à alagamentos e as estações climáticas

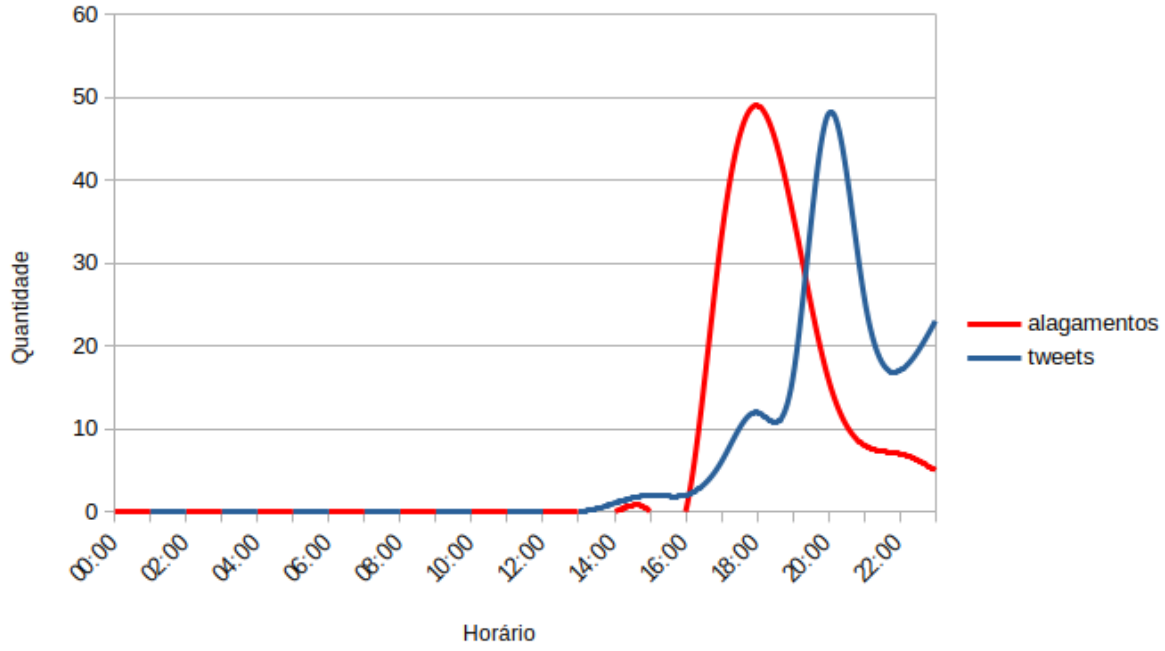


Fonte: Adaptada de Andrade *et al.* (2017).

Por último, na Figura 19 é apresentado uma comparação entre as séries temporais dos *tweets* geo referenciados relacionados com inundações e dos incidentes de alagamentos da cidade de São Paulo do dia 24 de fevereiro de 2017. Dessa forma, observa-se que nesse dia houve 154 ocorrências de alagamentos registradas de maneira horária pelo CGE-SP e 153 mensagens publicadas no *Twitter* que possuem relacionamento com alagamentos. Inclusive, nota-se na Figura 19 que os *tweets* ocorrem frequentemente após a incidência dos desastres naturais, pois

o pico das mensagens disseminadas no *Twitter* acontece após ao pico dos acontecimentos de desastres naturais. Portanto, esta característica de relacionamento entre as séries temporais dos *tweets* e das ocorrências de alagamentos, proporciona que as mensagens publicadas no *Twitter* sejam empregadas para auxiliar a etapa de resposta da GD, visto que essa fase tem o objetivo de promover operações de busca e salvamento da população afetada pelas inundações (POSER; DRANSCH, 2010).

Figura 19 – Comparação entre as séries temporais das mensagens publicadas no *Twitter* e as ocorrências de inundações do dia 24 de fevereiro de 2017 na cidade de São Paulo



Fonte: Elaborada pelo autor.

4.1.2 Bases de Dados

Esta etapa possui a função de extrair os dados de diferentes espaços informacionais com o intuito de proporcionar as informações necessárias para a confecção do "Conjunto Verdade" (Subseção 4.1.6). Dessa forma, os conjuntos de dados utilizados nesta abordagem estão descritos na Tabela 4, assim observa-se a categoria das informações, quais são os fornecedores dos dados, quais são as estratégias de captação das informações, a quantidade de dados coletados e o período das extrações. Logo após, são descritos detalhadamente as informações contidas na tabela.

Primeiramente, o *Twitter* é uma rede social que viabiliza a captação de inúmeras mensagens disseminadas em sua plataforma de forma gratuita via *Application Programming Interface* (API). Desse modo, para a captação das **mensagens publicadas no *Twitter*** ("data", "horário do compartilhamento", "texto", "coordenadas geográficas") foi efetuada uma parceria com o grupo de pesquisa *A Geospatial Open collaboRative Architecture for building resilience against*

Tabela 4 – Bases de Dados e métodos de extração de informações utilizados

Tipo de Informação	Fornecedor	Estratégia de Captação	Quantidade	Período
Mensagens publicadas no *Twitter*	*Twitter*	*Stremming API*	4,031 milhões	07/11/2016 até 06/11/2018
Dados Meteorológicos	INMET	Manual	33,576 mil	01/01/2015 até 30/10/2018
Ocorrências Históricas de Alagamentos	CGE-SP	*Web Scrapping*	4,904 mil	01/01/2015 até 30/10/2018

Fonte: Elaborada pelo autor.

disasters and extreme events (AGORA)[3] do Instituto de Ciências Matemáticas e de Computação (ICMC)[4] pertencente a Universidade de São Paulo (USP)[5], onde os membros atuantes desse grupo de pesquisa nos concederam 4,031 milhões de *tweets* geo referenciados do município de São Paulo empregados em alguns trabalhos científicos realizados por eles (por exemplo, Andrade *et al.* (2018), Restrepo-Estrada *et al.* (2018), Andrade *et al.* (2020)).

Dessa maneira, as mensagens disseminadas no *Twitter* do período de 7 de novembro de 2016 até 6 de novembro de 2018 foram coletadas pelos membros ativos do grupo de pesquisa AGORA com o auxílio de um *crawler*, ou seja, um mecanismo computacional especialista na extração de dados da *Web* (HEYDON; NAJORK, 1999). Inclusive, os pesquisadores do AGORA utilizaram algumas caixas delimitadores para a captação dos *tweets* geo localizados, assim foram coletadas somente informações presentes geograficamente nestas duas caixas que envolvem a cidade de São Paulo, sendo a primeira ao norte do município (-46.95, -23.62, -46.28, -23.33), já a segunda ao sul da capital paulista (-46.95, -23.91, -46.28, -23.62) (ANDRADE *et al.*, 2018).

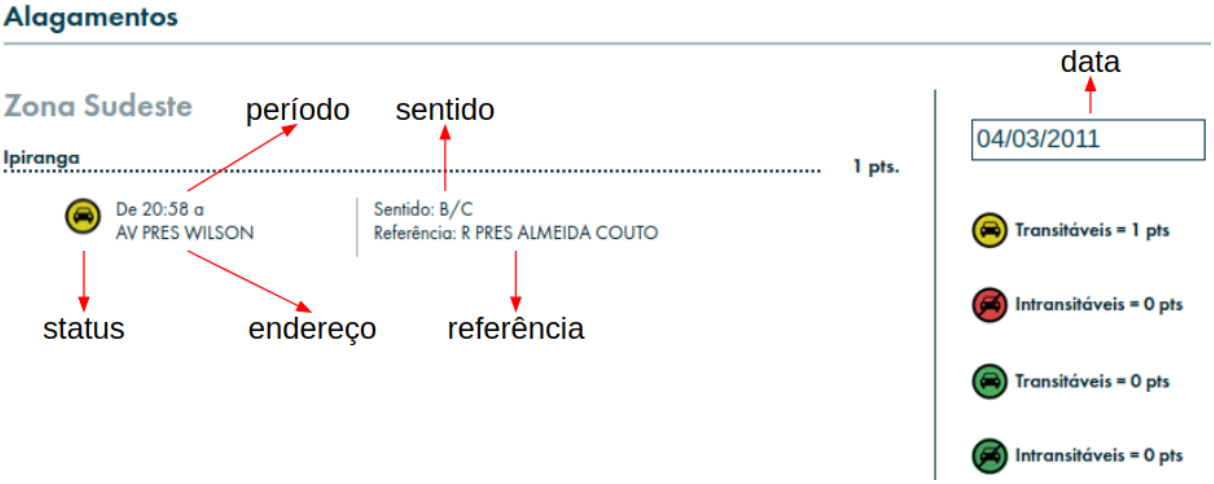
Ademais, a coleta dos **dados meteorológicos** do município de São Paulo foi realizada manualmente, pois as informações estão disponíveis publicamente na seção de informações meteorológicas históricas do portal *Web* do INMET, assim foram captadas 33.576 observações de medições automáticas efetuadas de hora em hora pelas estações climáticas do INMET do período de 1 de janeiro de 2015 até 30 de outubro de 2018. Aliás, os seguintes dados foram captados: “data de medição”, “horário de medição”, “temperatura”, “umidade relativa do ar”, “temperatura do ponto do orvalho”, “pressão atmosférica” e “precipitação”.

Já para a extração das **ocorrências históricas de alagamentos** da cidade de São Paulo, foi desenvolvido um mecanismo computacional em *Python* capaz de extrair as informações de inundações do portal *Web* do CGE-SP. Assim, a seguir no Algoritmo 1 apresentam-se as etapas

[3] url: <http://www.agora.icmc.usp.br/site/language/en/>
[4] url: <https://www.icmc.usp.br/>
[5] url: <https://www5.usp.br/>

executadas pelo *crawler* e na Figura 20 consta os elementos captados da página *Web* do CGE-SP pelo mecanismo de extração. Inclusive, foi necessário o desenvolvimento de um mecanismo computacional capaz de coletar as informações da página *Web* do CGE-SP, pois o CGE-SP não possui uma API para o fornecimento de seus dados públicos de incidentes de alagamentos do município de São Paulo, além disso, leva-se um tempo exacerbado para a execução da tarefa de captar todas as informações do período de 1 de janeiro de 2015 até 30 de outubro de 2018 manualmente.

Figura 20 – Exemplo das características relacionadas com ocorrências de alagamentos captadas da página *Web* do CGE-SP



Fonte: Elaborada pelo autor.

Desse modo, nota-se no Algoritmo 1 que foi utilizado para auxiliar o processo de extração de dados da página *Web* do CGE-SP a biblioteca escrita em *Python* chamada *Beautiful Soup*[6], no qual possui a função de facilitar o processo de extração de dados do HTML das páginas *Web*. Aliás, o seguinte endereço eletrônico corresponde a "url_base" contida no Algoritmo 1: "https://www.cgesp.org/v3/alagamentos.jsp?dataBusca='+dia+'/'+mes+'/'+ano+'+&enviaBusca =Buscar", onde as variáveis "dia", "mes" e "ano" equivalem ao procedimento contido na linha 5 do algoritmo.

Por último, o *crawler* desenvolvido para captar os dados do CGE-SP navega pelos elementos contidos no HTML da página *Web* e coleta 4904 ocorrências de alagamentos do período de 1 de janeiro de 2015 até 30 de outubro de 2018, no qual contém as seguintes informações que também podem ser observadas na Figura 20: "data", "período", "endereço", "referência", "sentido" e "status". Posteriormente, esses dados são persistidos em um banco de dados MongoDB[7]. Inclusive, o banco de dados não relacional chamado MongoDB foi empregado como estratégia de armazenamento de dados desta etapa, pois segundo Aghi *et al.* (2015) esta abordagem de armazenamento possui um desempenho superior na tarefa de realizar consultas

[6] url: <https://pypi.org/project/beautifulsoup4/>
[7] url: <https://www.mongodb.com/>

Algoritmo 1 – Captação de dados do portal *Web* do CGE-SP

```
1: procedimento CRAWLER(data_inicial, data_final) ▷ Informar o período de captação dos dados
2:     Criar um Dataframe para salvar as ocorrências de alagamentos encontradas
3:     para todos os dias presentes no período de captação de dados faça
4:         Dividir os dados referentes ao dia, mês e ano da data que está sendo analisada
5:         Definir a url base que será navegada pelo crawler
6:         Inserir os dados referentes ao dia, mês e ano na url base
7:         Inicializar um objeto Beautiful Soup com a url base
8:         Encontrar todos os itens da classe "tb-ponto-de-alagamento" na página Web
9:         Verificar a quantidade de itens descobertos da classe "tb-ponto-de-alagamento"
10:        se algum item da categoria "tb-ponto-de-alagamento" for descoberto então
11:            para todos os itens descobertos da classe "tb-ponto-de-alagamento" faça
12:                Encontrar todos os itens da classe "ponto-de-alagamento"
13:                para todos os itens descobertos da classe "ponto-de-alagamento" faça
14:                    se achar itens da classe "ativo-transitavel" ou "inativo-transitavel" então
15:                        Persistir o status como "transitável" no Dataframe
16:                    senão
17:                        Persistir o status como "intransitável" no Dataframe
18:                    fim se
19:                    Encontrar todos os itens da classe "arial-descr-alag"
20:                    cont_end_ref ← 0
21:                    para todos elementos encontrados da classe "arial-descr-alag" faça
22:                        Achar todos os itens com "text=True"
23:                        cont_features ← 0
24:                        para todos itens "text=True" descobertos faça
25:                            se cont_end_ref == 0 e cont_features == 0 então
26:                                Salvar o item na coluna "período" do Dataframe
27:                                se cont_end_ref == 1 e cont_features == 0 então
28:                                    Salvar o item na coluna "endereço" do Dataframe
29:                                    se cont_end_ref == 1 e cont_features == 0 então
30:                                        Salvar o item na coluna "sentido" do Dataframe
31:                                    senão
32:                                        Salvar o item na coluna "referência" do Dataframe
33:                                    fim se
34:                                fim se
35:                            fim se
36:                            cont_features ← +1
37:                        fim para
38:                        cont_end_ref ← +1
39:                    fim para
40:                fim para
41:            fim para
42:        fim se
43:    fim para
44:    Persistir as informações contidas no Dataframe no banco de dados MongoDB
45: fim procedimento
```

complexas e inserir grandes volumes de informações quando comparado com os tradicionais bancos de dados relacionais (por exemplo, MySQL[8], PostgreSQL[9], entre outros).

4.1.3 Seleção dos Dados

Esta fase da abordagem possui a função de escolher as informações multimodais essenciais para a criação de um espaço amostral coeso, no qual posteriormente será empregado como pré-requisito da etapa de Engenharia de Características (Subseção 4.1.5).

Primeiramente, para a seleção dos **dados textuais** captados do *Twitter*, realizou-se o processo de filtragem textual baseado em um conjunto de palavras-chave escritas na Língua Portuguesa relacionadas com fenômenos naturais (por exemplo, "chuva", "alagamento" e "inundação"), inclusive notam-se essas palavras na Tabela 5. Aliás, as palavras-chave utilizadas nesta etapa foram escolhidas com base em algumas pesquisas encontradas na literatura, no qual analisam os dados textuais de redes sociais escritos na Língua Portuguesa para contribuir com a GD (por exemplo, Assis *et al.* (2016), Andrade *et al.* (2017), Andrade *et al.* (2018), Andrade (2020)).

Tabela 5 – Palavras-chave relacionadas com fenômenos naturais

alagamento, alagado, alagada, alagando, alagou, alagar, chove, chova, chovia, chuva, chuvarada, chuvosa, chuvoso, chuvona, chuvinha, chuvisco, chovendo, dilúvio, enchente, enxurrada, garoa, inundação, inundada, inundado, inundar, inundam, inundou, temporal, temporais, tromba d'água

Fonte: Adaptada de Andrade *et al.* (2017).

Desse modo, para a efetuação da etapa de filtragem textual dos *tweets*, foi confeccionado nesta pesquisa um mecanismo computacional escrito em *Python* capaz de realizar buscas do tipo *substring* e efetuar verificações geográficas, ou seja, selecionar mensagens publicadas no *Twitter* que possuam ao menos uma das palavras-chave da Tabela 5 e que estejam localizadas geograficamente na cidade de São Paulo. Dessa maneira, obtivemos 13.802 *tweets*, no qual representam 0,34% de 4,031 milhões de mensagens captadas na seção de Bases de Dados (Subseção 4.1.2). Inclusive, essa porcentagem pequena de *tweets* geo localizados relacionados com inundações também foi observada em outros trabalhos encontradas na literatura que utilizam redes sociais para auxiliar a GD (por exemplo, Huang e Xiao (2015), Albuquerque *et al.* (2015)).

Além disso, quanto aos **dados históricos de inundações** somente foram selecionadas as observações que continham "data", "horário", "endereço", "sentido", "referência" e "*status*", ou seja, todas as características possíveis informadas pelo CGE-SP. Assim, para a confecção deste espaço amostral foram excluídas observações que continham dados faltantes, visto que a execução de técnicas de ML em espaços amostrais com informações ausentes pode exibir erros de execução (FACELI *et al.*, 2011). Dessa forma, as características "endereço" e "referência"

[8] url: <https://www.mysql.com/>
[9] url: <https://www.postgresql.org/>

são importantes, pois podem ser utilizadas em *softwares* de inferência de geo localização (por exemplo, *Google Geocoding API*[10]) para a obtenção das coordenadas geográficas das ocorrências de alagamentos. Já as informações "data" e "horário", são indispensáveis para o processo de fusão espaço-temporal contido na seção de criação do "Conjunto Verdade" (Subseção 4.1.6).

Ademais, foram selecionadas as **informações meteorológicas** pertencentes ao período de 7 de novembro de 2016 até 30 de outubro de 2018, pois esse intervalo temporal corresponde ao resultado da intersecção entre as datas de publicação dos *tweets* e os dados climáticos. Inclusive, as informações referentes a "data" e o "horário" das ocorrências de alagamentos são obrigatórias para o procedimento de fusão espaço-temporal pertencente a seção de criação do "Conjunto Verdade" (Subseção 4.1.6). Aliás, diferente dos dados históricos de alagamentos, as informações meteorológicas podem conter algumas observações com dados ausentes, pois essas observações não precisam estarem completas para serem utilizadas por técnicas de ML, visto que existe a possibilidade da criação de um mecanismo computacional capaz de estimar os valores das informações ausentes (por exemplo, aplicação da técnica de interpolação linear) (FACELI *et al.*, 2011).

Além disso, seleciona-se a característica climática chamada "precipitação", pois ela é imprescindível para o modelo de Fusão Multimodal, já que quanto mais intensa é essa característica meteorológica, consequentemente é a quantidade de água que chega de forma simultânea nos rios, desse modo o surgimento dos alagamentos ocorre devido ao excesso do volume hídrico nos sistemas de drenagem das regiões urbanas (TUCCI; BERTONI *et al.*, 2003). Ademais, a "pressão atmosférica" é importante para esta abordagem, porque ela é inversamente proporcional a precipitação, ou seja, períodos com grandes volumes de chuva possuem baixo nível de pressão atmosférica (MAKARIEVA *et al.*, 2014).

Aliás, a informação meteorológica chamada "temperatura" apresenta-se como uma característica crucial para o modelo de Fusão Multimodal, já que ela é diretamente proporcional a precipitação, além de que o aumento da temperatura está relacionado com a absorção de uma parcela da energia solar pelas superfícies impermeáveis (por exemplo, asfalto e concreto) (TUCCI; BERTONI *et al.*, 2003). Já a "umidade relativa do ar" é imprescindível para esta abordagem, porque ela indica a probabilidade da formação de nuvens de chuva, visto que o processo de condensação é iniciado quando a umidade relativa do ar alcançar 100% de sua capacidade (MARTINS, 2006). Por último, a "temperatura do ponto do orvalho" é importante para esta abordagem, pois de acordo com Talaia e Vigário (2016) ela sinaliza a temperatura que o ar úmido se resfria, além de indicar a quantidade de água presente em uma parcela do ar.

4.1.4 Classificação Manual das Informações Textuais

Esta etapa da abordagem possui o objetivo de classificar manualmente as informações textuais advindas do *Twitter*, visto que para a aplicação das técnicas de ML que serão utilizadas na

[10] url: <https://developers.google.com/maps/documentation/geocoding/overview>

criação do modelo de Fusão Multimodal as observações do espaço amostral necessitam de rótulos. Inclusive, esta metodologia de classificação manual foi baseada em pesquisas encontradas na literatura, no qual possuem o objetivo de minerar dados de redes sociais para obter a SAW de desastres naturais e auxiliar a GD (por exemplo, Bruijn *et al.* (2020)).

Primeiramente, as informações textuais filtradas na fase de Seleção dos Dados (Subseção 4.1.3) foram transmitidas para esta etapa com o intuito de serem classificadas por três juízes, no qual foram responsáveis pela classificação manual de **13.802 *tweets*** quanto a presença de relacionamento com inundações ou chuva. Desse modo, foram categorizados **4.527 (33%) *tweets* como "relevantes"** e **9.275 (67%) como "irrelevantes"**. Aliás, no decorrer do processo de classificação manual das mensagens captadas do *Twitter*, os responsáveis pela classificação encontraram diversas mensagens escritas informalmente, inclusive uma parcela dos *tweets* classificados como "irrelevantes" perante a presença de conteúdos relacionados com chuva ou inundações apresentaram conteúdos metafóricos, por exemplo: "bolinho de chuva", "chuva de bençãos", "enxurrada de lágrimas", "inundação de ideias", "pensamento atemporal", entre outros. Dessa forma, na Tabela 6 nota-se alguns exemplos de mensagens publicadas no *Twitter* rotuladas como "relevantes" e "irrelevantes" pelos juízes deste processo de classificação manual.

Tabela 6 – Exemplos de *tweets* classificados como "relevantes" e "irrelevantes" para o contexto de chuvas e alagamentos

Data	Horário	Tweet	Classificação
01/02/2017	17:42:22	Carro danificado por enchente	Relevante
03/01/2017	23:57:27	Bolinho de chuva da sogra	Irrelevante
20/03/2018	21:00:52	Meu Deus do céu como eu vou sair daqui? #sp #alagamento #socorro	Relevante
09/08/2017	14:38:29	Hoje chorei mais que chuva em temporal	Irrelevante

Fonte: Elaborada pelo autor.

Por último, para a avaliação do grau de confiança das informações classificadas manualmente pelos três juízes, então submetemos os rótulos dos *tweets* a um mecanismo computacional capaz de calcular o coeficiente *Alpha Krippendorf*, no qual é uma estratégia de avaliação estatística capaz de estimar o grau de concordância entre dois ou mais avaliadores (KRIPPENDORFF, 2004). Dessa maneira, alcançamos um **grau de concordância de 0,66** entre os avaliadores, inclusive este resultado demonstra um nível de **concordância substancial** entre os juízes, porque o grau de anuência perfeito é obtido com o resultado mais próximo de um (1), já o nível de anuência imperfeito ou ausente é demonstrado com a resposta do coeficiente *Alpha Krippendorf* mais próxima de zero (0).

4.1.5 Engenharia de Características

A etapa de Engenharia de Características possui o objetivo de pré-processar as informações resultantes da etapa de Seleção dos dados (Subseção 4.1.3) e proporcionar mais qualidade

as informações multimodais que serão utilizadas em etapas posteriores, ou seja, possibilitar a confecção de dados ausentes de ruídos, balanceados, com dimensões semelhantes e convertidos para caracteres que os algoritmos de ML consigam interpretar.

4.1.5.1 Informações Textuais

Primeiramente, esta fase da Engenharia de Características possui a tarefa de pré-processar as informações textuais captadas do *Twitter*. Desse modo, aplicaram-se os procedimentos de "limpeza de dados" e "transformação de dados" inerentes a metodologia de MT. A princípio, a fase de limpeza de dados é responsável pela diminuição da quantidade de ruídos presentes no conjunto amostral de *tweets* (por exemplo, *emoticons*, abreviações, endereços eletrônicos, entre outros). Já a fase de transformação de dados, atribui-se a tarefa de converter os dados textuais de caracteres simbólicos para numéricos.

Dessa maneira, com o intuito de executar a fase de "limpeza de dados", foi desenvolvido para esta pesquisa um algoritmo capaz de remover ruídos textuais das mensagens disseminadas no Twitter. Inclusive, o algoritmo criado para a execução desta fase pode ser observado no Algoritmo 2, além disso, para que o mecanismo computacional possa realizar seus objetivos magistralmente, portanto empregamos as seguintes bibliotecas escritas em *Python*: *Natural Language Toolkit* (NLTK)[11], *CyHunspell*[12], *Re*[13] e *Gensim*[14], no qual possuem o intuito de auxiliarem no PLN dos *tweets*. Além disso, o algoritmo de pré-processamento textual desenvolvido para esta etapa da pesquisa apresenta as seguintes funcionalidades:

1. Remoção de endereços eletrônicos, perfis, *stop words*, caracteres duplicados e especiais dos *tweets*;

2. Transformação dos *tweets* em *tokens*;

3. Substituição das *hashtags* contidas nos *tweets* pelas respectivas traduções;

4. Correção das palavras informais dos *tweets* por palavras escritas na norma culta da Língua Portuguesa.

Dado o exposto, o Algoritmo 2 utiliza os seguintes glossários para auxiliar o PLN das mensagens publicadas no *Twitter*: dicionário de *hashtags*, dicionário de palavras-chave coloquiais e dicionário da Língua Portuguesa. Desse modo, os dois primeiros glossários foram confeccionados manualmente a partir da análise das *hashtags* e das sentenças coloquiais mais

[11] url: <https://www.nltk.org/>
[12] url: <https://pypi.org/project/cyhunspell/>
[13] url: <https://pypi.org/project/re2/>
[14] url: <https://pypi.org/project/gensim/>

Algoritmo 2 – Processo de limpeza de dados dos *tweets*

para todas observações contidas no espaço amostral de *tweets* **faça**
    Remover os endereços eletrônicos com o auxílio de expressões regulares
    Remover os perfis e caracteres duplicados de *tweets* com o apoio do método *TweetTokenizer* presente na biblioteca NLTK
    Criar um vetor para salvar os *tokens* corrigidos pelo dicionário de *hashtags*
    para todos os *tokens* existentes nas mensagens publicadas no *Twitter* **faça**
        para todos os itens presentes no dicionário de *hashtags* **faça**
            se algum *token* pertença ao Apêndice A **então**
                Substituir o *token* pelo significado da *hashtag*
                Salvar a correção do *token* no vetor
            fim se
        fim para
    fim para
    Remover os caracteres especiais das mensagens publicadas no *Twitter* com o auxílio de expressões regulares
    Remover as *stop words* dos *tweets* com o auxílio da biblioteca NLTK
    Criar um vetor para persistir as correções sugeridas pelo *LibreOffice*
    para para todos os *tokens* das mensagens disseminadas no *Twitter* **faça**
        se o *token* está escrito na norma culta da Língua Portuguesa conforme as *Word Embeddings* propostas por Hartmann *et al.* (2017) **então**
            Salvar o *token* no vetor de correções do *LibreOffice*
        senão
            se o *token* está contido no Apêndice B **então**
                Substituir o *token* pela respectiva palavra escrita na norma culta da Língua Portuguesa
                Persistir a correção do *token* no vetor de correções do *LibreOffice*
            senão
                se a extensão do dicionário da Língua Portuguesa pertencente ao *LibreOffice* sugira correções para o *token* analisado **então**
                    para todas as sugestões de correção do *LibreOffice* **faça**
                        se a correção está contida no espaço amostral de *Word Embeddings* da Língua Portuguesa proposto Hartmann *et al.* (2017) **então**
                            Salvar a correção do *token* no vetor de correções do *LibreOffice*
                            Finalizar a sequência de repetições de sugestões de correção
                        fim se
                    fim para
                fim se
            fim se
        fim se
    fim para
    Persistir os vetores com as correções em uma estrutura de dados
    Salvar a estrutura de dados no MongoDB
fim para

frequentes presentes nos *tweets* relacionados com alagamentos, inclusive no Apêndice A observa-se o glossário de *hashtags* e seus respectivos significados (por exemplo, "#saopaulo4you" traduz-se para "A cidade de São Paulo para você"), já no Apêndice B nota-se o glossário de palavras-chave coloquiais e suas respectivas traduções para a norma culta da Língua Portuguesa (por exemplo, "trampando" traduz-se para "trabalhando"). Aliás, o terceiro glossário corresponde a extensão do dicionário da Língua Portuguesa utilizado pelo *LibreOffice*, onde foi empregado pelo Algoritmo 2 com o auxílio da biblioteca *CyHunspell* para a obtenção de sugestões de correções das palavras errôneas.

Além disso, observa-se no Algoritmo 2 que os endereços eletrônicos e os caracteres especiais foram removidos das mensagens publicadas no *Twitter* com o auxílio de expressões regulares, assim a biblioteca que auxilia a aplicação dessas expressões no ambiente de programação *Python* é a *Re*. Já o modelo de *Word Embeddings* utilizado pelo algoritmo de pré-processamento textual, possui o objetivo de possibilitar a verificação da existência dos elementos textuais na norma culta da Língua Portuguesa, visto que esse modelo foi treinado pelo Núcleo Interinstitucional de Linguística Computacional (NILC) do ICMC-USP em um gigante *corpus* textual da Língua Portuguesa (por exemplo, informações extraídas da Wikipédia[15] e outros ambientes informacionais digitais), aliás a biblioteca *Gensim* que garante o carregamento do modelo de *Word Embeddings* criado pelo NILC para o ambiente de programação *Python* utilizado neste trabalho. Ademais, a biblioteca NLTK que é responsável por auxiliar os processos de conversão das mensagens para *tokens* e remoção de *stop words* da Língua Portuguesa (por exemplo, "seu", "com", "para", entre outras) dos *tweets*.

Por último, as informações resultantes da fase de limpeza de dados são conduzidas para a fase de transformação de dados com o intuito de converter os elementos textuais para numéricos, visto que as máquinas possuem dificuldade para interpretar informações simbólicas. Desse modo, para a execução da fase de transformação de dados, foram utilizadas algumas estratégias encontradas em pesquisas da literatura que aplicam a etapa de pré-processamento da MT em *tweets* (por exemplo, Aguiar *et al.* (2018), Bruijn *et al.* (2020)). Dessa maneira, as estratégias escolhidas são: BOW, TF-IDF e *Word Embeddings* treinadas por Hartmann *et al.* (2017) dos tipos Word2Vec e FastText das categorias Skip-gram e CBOW com 50 e 100 dimensões.

4.1.5.2 Dados históricos de alagamentos

Esta fase da Engenharia de Características possui o objetivo de pré-processar as ocorrências de alagamentos da cidade de São Paulo, além de proporcionar informações sobre inundações ausentes de inconsistências e que possam ser interpretadas por algoritmos de ML. Dessa forma, para a execução desta etapa aplicaram-se os processos de "limpeza de dados", "transformação de dados textuais para coordenadas geográficas" e "agrupamento de informações".

A princípio, o processo de "limpeza de dados" desta etapa tem o objetivo de remover

[15] url: <https://pt.wikipedia.org/>

os ruídos dos dados históricos de alagamentos da cidade de São Paulo, pois as ocorrências de inundações contidas no portal *Web* do CGE-SP apresentam diversos endereços escritos de maneira abreviada e informal (por exemplo, "es" é equivalente a "estrada" e "velha fepasa" corresponde ao endereço "comunidade hungara"). Inclusive, essas informações inconsistentes dificultam a captação das localizações dos desastres naturais por *softwares* capazes de inferir as coordenadas geográficas de eventos a partir do endereço.

Assim, para atender os objetivos da fase de "limpeza de dados" desta etapa da Engenharia de Características, foi criado um mecanismo computacional escrito em *Python* capaz de remover os ruídos das ocorrências de enchentes e proporcionar dados menos complexos para os *softwares* de indução de coordenadas geográficas (por exemplo, *Google Geocoding API*). Aliás, o mecanismo computacional elaborado nesta pesquisa pode ser observado no Algoritmo 3, além disso, a biblioteca *Pandas*[16] escrita em *Python* é empregada para auxiliar o mecanismo computacional na realização de suas tarefas de modo eficaz, pois ela proporciona estruturas de dados ágeis e flexíveis (por exemplo, *Dataframe*), além de métodos que facilitam a manipulação dos dados dessas estruturas (MCKINNEY, 2020).

Dado o exposto, o Algoritmo 3 utiliza dois glossários de dados informais com o intuito de corrigir ou remover palavras errôneas, inclusive esses dicionários foram criados manualmente a partir da análise das ocorrências históricas de inundações reportadas pelo CGE-SP. Desse modo, o primeiro pode ser notado no Apêndice C e representa as palavras-chave e abreviações mais frequentes presentes nas amostras do conjunto de dados e suas respectivas traduções para a norma culta da Língua Portuguesa. Já o segundo, pode ser observado no Apêndice D e retrata as expressões coloquiais mais assíduas do espaço amostral, inclusive essas informações são passíveis ao processo de remoção, pois a inclusão desses dados não impacta no desempenho do modelo de Fusão Multimodal, além de proporcionar um grau de complexidade maior para a interpretação dos endereços das ocorrências pelos *softwares* de inferência de coordenadas geográficas.

Logo após os dados serem processados pelo Algoritmo 3, as informações resultantes são enviadas para a fase de "transformação de dados textuais para coordenadas geográficas", onde possui o objetivo de converter os endereços dos desastres naturais para localizações geográficas no globo terrestre. Inclusive, nesta etapa foi utilizado o mecanismo computacional chamado *Google Geocoding API*, no qual é um *software* de PLN e dedução geográfica capaz de interpretar endereços e retornar as coordenadas geográficas. Dessa forma, a seguir na Tabela 7 notam-se alguns exemplos de incidências históricas de inundações e as respectivas datas, períodos de acontecimento e localizações geográficas dos fenômenos.

Atualmente, não existe uma ferramenta que possibilite a obtenção das coordenadas geográficas das regiões propícias ao acontecimento de inundações da cidade de São Paulo, desse modo conforme a teoria de Tobler (1970) que propicia a criação da hipótese de que mensagens

[16] url: <https://pypi.org/project/pandas/>

Algoritmo 3 – Processo de limpeza de dados das ocorrências históricas de alagamentos

procedimento PREPROCESS_ALAG(*dataframe_alag*) ▷ Informar os dados das ocorrências de inundações
    Criar uma estrutura de dados para salvar os dados pré-processados
    para todas observações do *Dataframe* de informações de alagamentos **faça**
        se o *status* corresponder a "transitável" **então**
            Salvar no *Dataframe* o tipo de alagamento como 1
        senão
            Salvar no *Dataframe* o tipo de alagamento como 0
        fim se
        Dividir o período das ocorrências de inundações em inicial e final de maneira horária
        se o endereço ou a referência possuir algum item do Apêndice D **então**
            Remover o elemento textual da observação
        fim se
        se o endereço ou a referência possuir algum item do Apêndice C **então**
            Substituir o elemento textual da observação pela respectiva tradução
        fim se
        se a referência for numérica **então**
            Concatenar o número da referência ao endereço com uma "," ▷ Exemplo de resultado: AVENIDA GUILHERME COTCHING, 16, São Paulo, Brasil
        senão
            Concatenar a referência ao endereço com a palavra "com" ▷ Exemplo de resultado: RUA CAPACHOS COM RUA ANDRÉ DE GUIMARÃES, São Paulo, Brasil
        fim se
        Salvar as informações pré-processadas em uma estrutura de dados
    fim para
    Salvar os dados da estrutura de dados no MongoDB
fim procedimento

Tabela 7 – Exemplos de endereço e coordenadas geográficas de casos de alagamento da cidade de São Paulo

Data	Período	Endereço	Latitude	Longitude
17/10/2019	De 17:15 até 18:40	Avenida das Pontes com Rua Danças Húngaras, São Paulo, Brasil	-23.480463	-46.3786941
28/10/2019	De 04:05 até 10:08	Rua Miguel Yunes com Avenida Interlagos, São Paulo, Brasil	-23.6841075	-46.691945

Fonte: Elaborada pelo autor.

publicadas no *Twitter* relacionadas com alagamentos estão próximas dos fenômenos. Então, necessita-se agrupar as informações históricas de alagamentos da capital paulista para determinar as áreas favoráveis de ocorrências inundações, visto que possíveis vítimas de fenômenos naturais tendem estar localizadas geograficamente dentro dessas áreas.

Dado o exposto, as informações resultantes da etapa de "transformação de dados textuais para coordenadas geográficas" são enviadas para a fase de "agrupamento de informações", no qual é responsável pela clusterização das coordenadas geográficas das ocorrências históricas de

alagamentos da cidade de São Paulo com o intuito de gerar as áreas propícias de acontecimento de inundações. Dessa forma, para cumprir essa tarefa, nesta fase são treinados e testados os seguintes algoritmos: DBSCAN, OPTICS e *Agglomerative clustering*, já que essas abordagens de agrupamento criam *clusters* de maneira volátil. Aliás, também foi comparado o desempenho dos algoritmos de agrupamento quanto a aplicação das seguintes estratégias de definição de distância de formação de grupos: empírica (variação de 100 metros até 5 quilômetros) e geoestatística (Semivariograma). Inclusive, a biblioteca *Scikit-Learn*[17](versão 0.24.0) foi a responsável por auxiliar a aplicação das técnicas de ML no ambiente de programação *Python*.

4.1.5.3 Informações Meteorológicas

O conjunto de informações meteorológicas apresenta algumas observações com atributos ausentes, assim necessita-se estimar os valores faltantes para não prejudicar a execução das abordagens de ML, portanto é possível notar na literatura que algumas pesquisas para inferir dados meteorológicos ausentes empregam técnicas de interpolação (por exemplo, Xu *et al.* (2013)). Inclusive, a interpolação é uma abordagem que a partir de fórmulas matemáticas realiza a estimativa dos valores medianos do espaço amostral (SHRYOCK; SIEGEL; LARMON, 1973). Já a Interpolação Linear, é uma técnica que utiliza polinômios lineares para elaborar novos valores dentro de um intervalo de um conjunto informacional conhecido (MEIJERING, 2002). Dessa forma, foi desenvolvido um *software* escrito em *Python* capaz de inferir os dados faltantes através da Interpolação Linear das diversas observações do espaço amostral de informações meteorológicas.

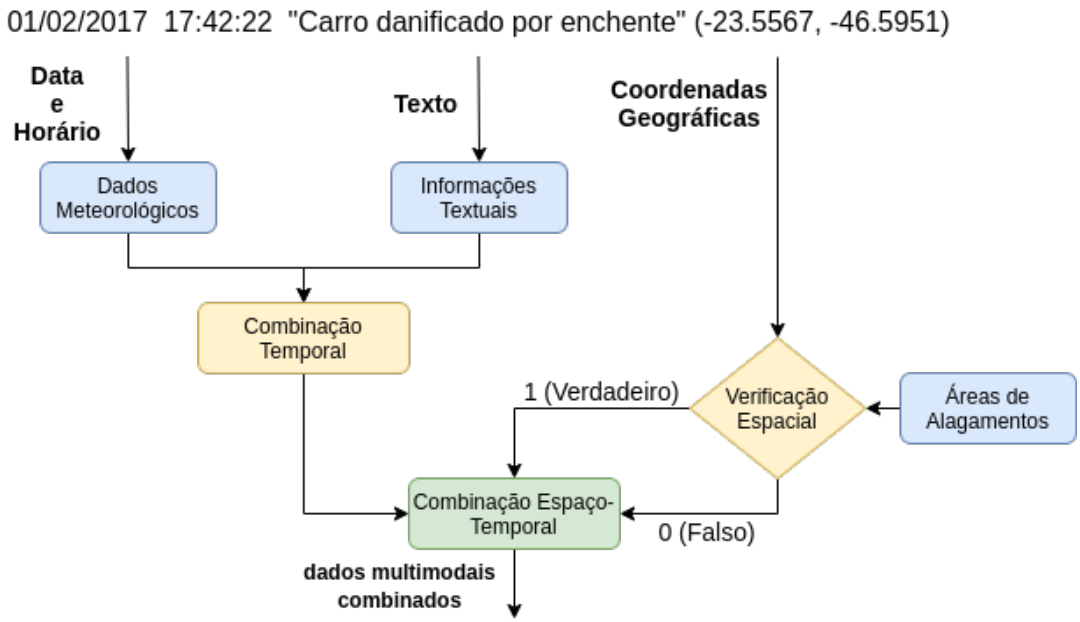
4.1.6 Criação do Conjunto Verdade

Esta etapa possui o objetivo de combinar de forma espaço-temporal as informações heterogêneas processadas na fase de Engenharia de Características, além de produzir dados relevantes para a criação de um Conjunto Verdade capaz de ser empregado para o treinamento e teste dos algoritmos de ML utilizados nesta abordagem. Dessa forma, a seguir na Figura 21 observa-se a metodologia utilizada nesta dissertação para a combinação dos dados multimodais, no qual é baseada em algumas pesquisas encontradas na literatura (por exemplo, Shyu *et al.* (2005), Chen *et al.* (2006), Yang *et al.* (2017), Pouyanfar *et al.* (2019)).

Na Figura 21 nota-se que os dados meteorológicos pré-processados (Subsubseção 4.1.5.3) são combinados com as informações textuais pré-processadas (Subsubseção 4.1.5.1) de maneira temporal, ou seja, os *tweets* compartilhados no dia 1 de janeiro de 2017 às 17 horas e 42 minutos são agrupados com as informações meteorológicas captadas do INMET do mesmo dia e período. Posteriormente, as informações geográficas dos *tweets* são submetidas a um processo de verificação espacial com as áreas de alagamentos derivadas dos dados históricos de inundação da cidade de São Paulo (Subsubseção 4.1.5.2). Inclusive, esse processo baseia-se na investigação da

[17] url: <https://scikit-learn.org/stable/>

Figura 21 – Processo de criação do Conjunto Verdade



Fonte: Elaborada pelo autor.

incidência de *tweets* em áreas de alagamentos, ou seja, utiliza-se a fórmula matemática chamada *Haversine* para calcular a distância entre as mensagens compartilhadas no *Twitter* e os centroides das zonas de inundação, assim atribui-se um (1) a resposta, caso essa distância seja inferior ao raio dos agrupamentos, do contrário atribui-se zero (0). Por fim, os dados meteorológicos e textuais combinados de forma temporal são agrupados linearmente com o resultado do processo de verificação espacial, assim originando a combinação espaço-temporal.

Aliás, a fórmula matemática denominada *Haversine* possui o objetivo de proporcionar a descoberta da distância geográfica entre duas localizações presentes no globo terrestre, inclusive nessa fórmula leva-se em consideração o grau de curvatura do planeta Terra (WINARNO; HADIKURNIAWATI; ROSSO, 2017). Assim, nota-se na Equação 4.1, a fórmula *Haversine* que é empregada extensamente em mecanismos computacionais cujo objetivo é processar informações geográficas.

$$d = 2.r.arcsin\left(\sqrt{sin^2\left(\frac{lat2-lat1}{2}\right)+cos(lat1).cos(lat2).sin^2.\left(\frac{long2-long1}{2}\right)}\right) \quad (4.1)$$

Dado o exposto, as variáveis presentes na Equação 4.1 são as seguintes: "**d**" corresponde a distância entre às duas coordenadas geográficas; "**r**" é o raio do planeta terra (6.731 quilômetros); "**lat1**" e "**lon1**" representam respectivamente a latitude e a longitude da primeira localização geográfica; "**lat2**" e "**lon2**" correspondem respectivamente a latitude e a longitude da segunda localização geográfica analisada.

Ademais, para a criação do Conjunto Verdade, utilizam-se as informações multimodais combinadas de maneira espaço-temporal e geram-se rótulos positivos e negativos para esses dados. Assim, foi atribuído o rótulo positivo (1) para todas as informações multimodais combinadas que continham informações textuais relacionadas com "alagamentos" ou "chuva" e que possuíam *tweets* contidos geograficamente em áreas de alagamentos da cidade de São Paulo, além de possuir dados climáticos compatíveis aos horários das incidências das inundações conforme o CGE-SP. Aliás, foi atribuído o rótulo negativo (0) para todas as combinações contrárias possíveis (por exemplo, informações multimodais combinadas que possuem *tweets* que não estão presentes geograficamente em áreas de inundações e que as informações textuais são relacionados com "inundações" ou "chuva", além de possuir dados meteorológicos pertencentes a períodos com ocorrência de alagamentos na cidade de são Paulo conforme o CGE-SP).

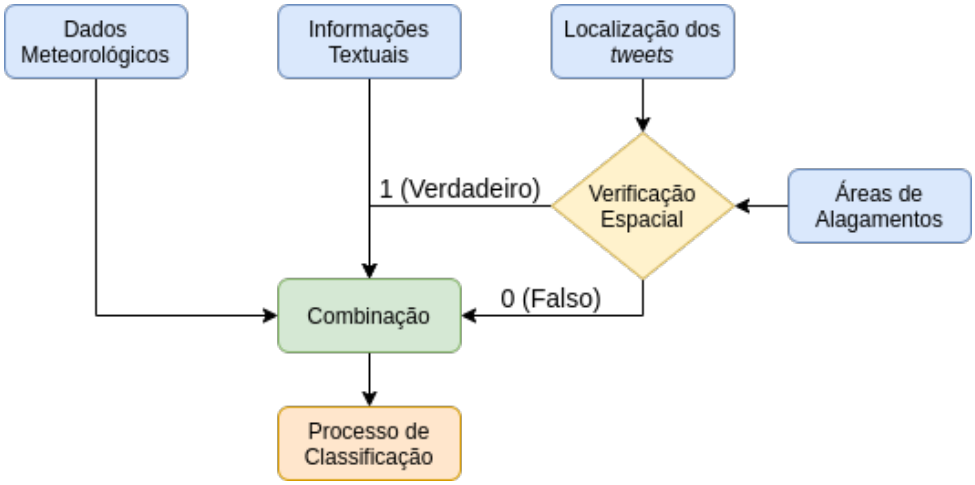
Por último, o Conjunto Verdade foi balanceado da seguinte maneira: 50% das observações positivas e 50% das observações negativas. Inclusive, esse procedimento de balanceamento de dados é importante, pois ao aplicar dados desbalanceados em algoritmos de ML há a possibilidade dos mecanismos computacionais favorecem as categorias majoritárias e terem o desempenho prejudicado (FACELI *et al.*, 2011). Ademais, somente 20% desta base de dados foi estabelecida como responsável pela validação da abordagem de Fusão Multimodal, assim os 80% restantes foram designados para a fase de treinamento dos modelos elaborados.

4.1.7 Modelos de Fusão Multimodal

Nesta etapa da abordagem, elabora-se quatro modelos de Fusão Multimodal que possuem o objetivo de obter a SAW de inundações da cidade de São Paulo e auxiliar a etapa de resposta da GD. Desse modo, os modelos de Fusão Multimodal elaborados e explorados nesta dissertação são os seguintes: Prévio; Tardio; Híbrido com enfoque na decisão das informações textuais; Híbrido com enfoque na decisão das informações meteorológicas. Inclusive, todos os modelos de fusão de informações multimodais desenvolvidos nesta dissertação foram embasados em informações encontradas em pesquisas da literatura (por exemplo, Atrey *et al.* (2010), Lopes (2015), Poria *et al.* (2016), Zadeh *et al.* (2017), Liu, Jiang e Zhao (2018), Bruijn *et al.* (2020)). Além disso, observa-se na Figura 26 o processo de classificação genérico empregado nos modelos de Fusão Multimodal, inclusive essa metodologia foi executada com auxílio da biblioteca *Scikit-Learn* (versão 0.24.0) escrita na linguagem de programação *Python*.

Dado o exposto, nota-se na Figura 22 que o processo de **Fusão Multimodal de modo prévio** elaborado nesta dissertação foi baseado na estratégia de combinação de informações multimodais de maneira prévia apresentada por Atrey *et al.* (2010). Assim, na Fusão Prévia há a combinação das informações textuais e meteorológicas ao nível de vetor de recursos, ou seja, combinam-se os dados textuais pré-processados e convertidos para caracteres numéricos com os dados dos sensores meteorológicos pré-processados. Posteriormente, agrupam-se essas informações multimodais com o resultado da verificação da incidência dos *tweets* em áreas de

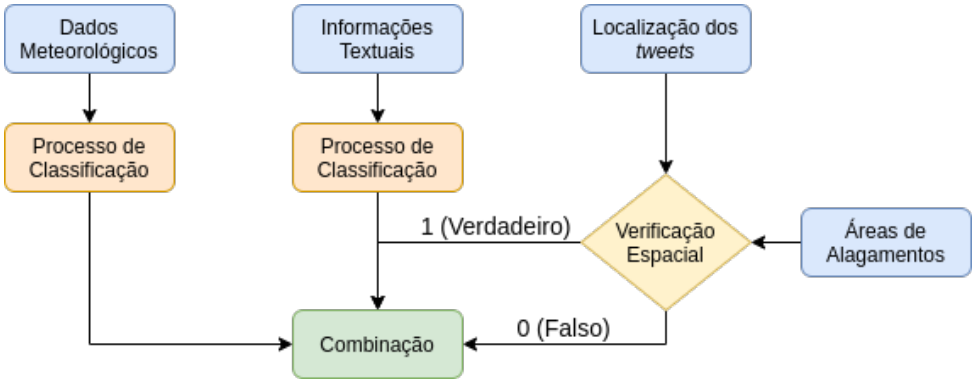
Figura 22 – Processo de Fusão Multimodal de modo prévio



Fonte: Elaborada pelo autor.

alagamentos da cidade de São Paulo. Por fim, essas informações agrupadas são submetidas ao modelo de classificação genérico detalhado na Figura 26, logo após avalia-se o desempenho desse mecanismo computacional responsável pela classificação das informações heterogêneas no conjunto de dados inédito separado na fase de criação do Conjunto Verdade (Subseção 4.1.6).

Figura 23 – Processo de Fusão Multimodal de modo tardio

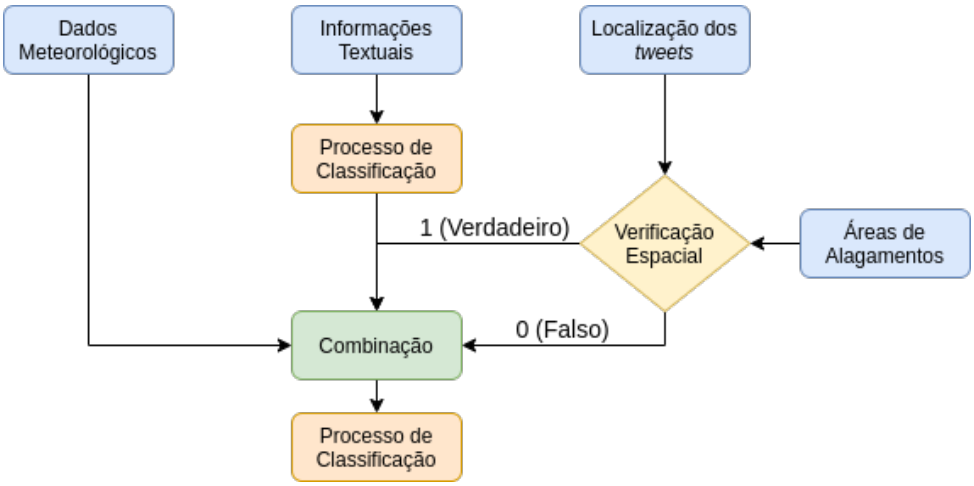


Fonte: Elaborada pelo autor.

Além disso, observa-se na Figura 23 o processo de **Fusão Multimodal de modo tardio**, no qual foi baseado na abordagem de combinação de informações midiáticas de maneira tardia proposta por Atrey *et al.* (2010). Assim, na Fusão Tardia existe a criação de dois modelos de classificação baseados na Figura 26, sendo o primeiro responsável pela classificação das informações textuais pré-processadas e convertidas para caracteres numéricos, já o segundo encarregado pela categorização dos dados meteorológicos pré-processados. Dessa forma, as informações resultantes dos modelos de classificação são combinadas linearmente com o resultado da verificação da ocorrência de *tweets* em áreas de inundações da cidade de São Paulo. Portanto, este modelo de Fusão Multimodal retorna resultados positivos (1), caso as mensagens publicadas

no *Twitter* sejam relacionados com “chuva” ou “inundação”, os *tweets* estejam contidos em áreas favoráveis ao acontecimento de alagamentos e os dados climáticos sejam propícios para a ocorrência desses fenômenos naturais, caso contrário o mecanismo computacional retorna resultados negativos (0). Inclusive, este modelo capaz de combinar informações heterogêneas de modo tardio é submetido a um processo de avaliação com o intuito de analisar seu desempenho em um conjunto de dados inédito.

Figura 24 – Processo de Fusão Multimodal de modo híbrido com enfoque na decisão das informações textuais

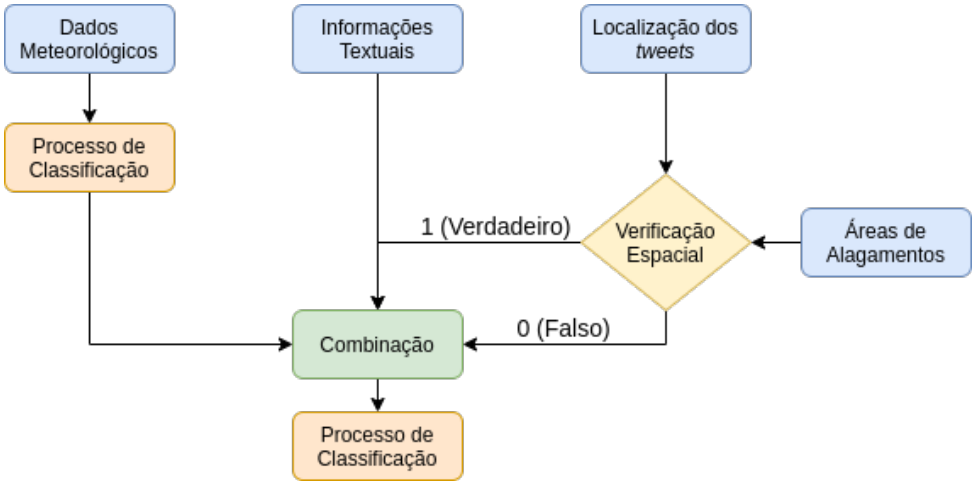


Fonte: Elaborada pelo autor.

Ademais, na Figura 24 nota-se o processo de **Fusão Multimodal de modo híbrido com enfoque na decisão das informações textuais**. Desse modo, para a elaboração dessa estratégia de combinação de informações midiáticas, primeiramente os dados textuais são submetidos a uma metodologia de classificação de dados genérica detalhada na Figura 26, no qual proporciona a elaboração de um modelo de identificação de *tweets* relacionados com “inundações” e “chuva”. Posteriormente, combinam-se as informações resultantes do modelo de classificação de *tweets* com os dados meteorológicos e o resultado da verificação da ocorrência dos *tweets* em áreas de alagamentos da cidade de São Paulo. Por fim, treina-se um mecanismo computacional baseado na metodologia descrita na Figura 26 com os dados ao nível de recurso e decisão agrupados, além disso, avalia-se o desempenho do modelo resultante quanto a aplicação em um conjunto de dados inédito criado na seção de criação do Conjunto Verdade (Subseção 4.1.6).

Além disso, na Figura 25 observa-se o processo de **Fusão Multimodal de modo híbrido com enfoque na decisão das informações meteorológicas**. Dessa forma, para a elaboração dessa estratégia de combinação de dados heterogêneos, primeiramente as informações meteorológicas são submetidas a uma metodologia de classificação de dados genérica detalhada na Figura 26, no qual proporciona a elaboração de um modelo de identificação de possíveis ocorrências de inundações da cidade de São Paulo. Posteriormente, combinam-se as informações resultantes do modelo de identificação de alagamentos com as informações textuais e o resultado

Figura 25 – Processo de Fusão Multimodal de modo híbrido com enfoque na decisão das informações meteorológicas



Fonte: Elaborada pelo autor.

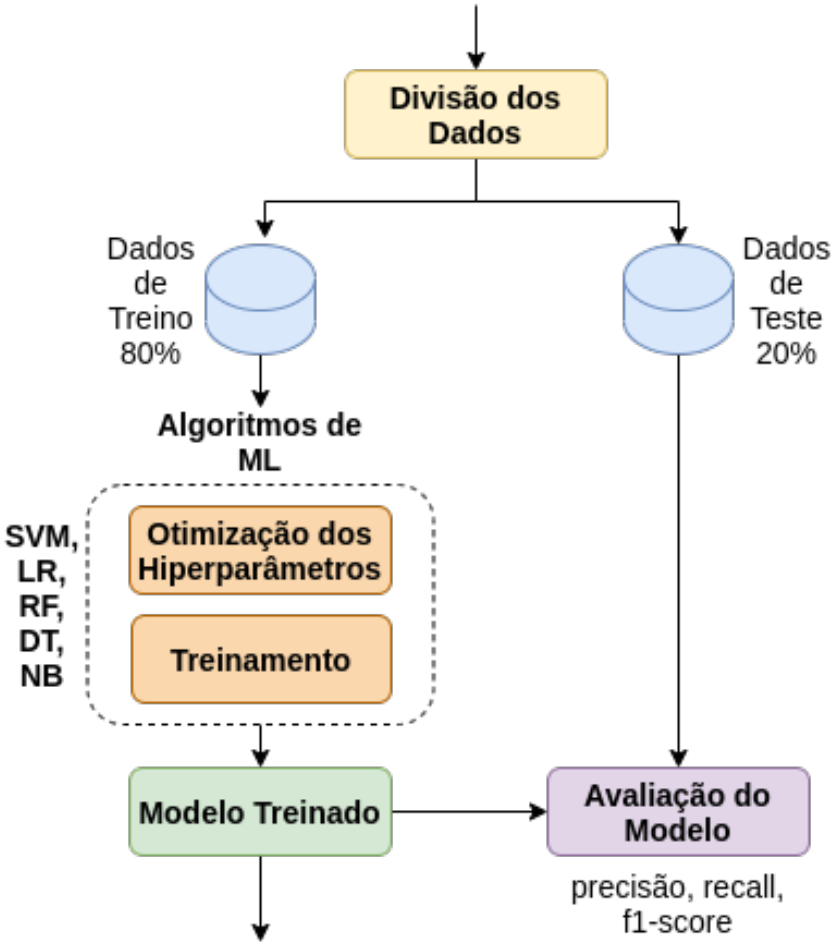
da verificação da ocorrência dos *tweets* em regiões de inundações da cidade de São Paulo. Por último, treina-se um modelo de classificação baseado na metodologia descrita na Figura 26 com os dados ao nível de recurso e decisão combinados, além disso, avalia-se o desempenho do mecanismo computacional resultante quanto a aplicação em um conjunto de dados inédito criado na seção de criação do Conjunto Verdade (Subseção 4.1.6).

Ademais, na Figura 26 nota-se o **processo de classificação genérico** empregado nos modelos de combinação de informações multimodais elaborados nesta pesquisa. Desse modo, observa-se que primeiramente os dados são divididos com o intuito de elaborar conjuntos informacionais responsáveis pelo treinamento e teste dos algoritmos de ML, assim esta fase foi organizada da seguinte maneira: 80% dos dados são reservados como conjunto de treinamento, já as 20% restantes são destinados como conjunto de avaliação. Inclusive, essa estratégia é adotada nesta abordagem, pois o conjunto de treinamento é empregado para a indução e ajuste dos modelos de ML, já o conjunto de teste possui a função de simular dados inéditos para medir o desempenho dos mecanismos computacionais treinados (FACELI *et al.*, 2011).

Posteriormente, os algoritmos de ML chamados SVM, LR, RF, DT e NB são submetidos ao processo de otimização de hiperparâmetros, ou seja, treina-se os algoritmos na base de dados de treinamento conforme um conjunto de parâmetros pré-definidos e aplica-se o processo de avaliação cruzada em todos os mecanismos computacionais de classificação treinados. Desse modo, nesta fase são escolhidos os hiperparâmetros que proporcionaram maior precisão aos algoritmos de ML, inclusive este procedimento é importante, pois o ajuste desses parâmetros propicia modelos de ML mais precisos (ROSSI, 2009).

Logo após a seleção dos hiperparâmetros que proporcionaram desempenhos mais interessantes aos mecanismos de classificação, treina-se os algoritmos de ML com os parâmetros selecionados anteriormente na base de dados de treinamento. Inclusive, os algoritmos de ML

Figura 26 – Processo de classificação genérico



Fonte: Elaborada pelo autor.

utilizados nesta abordagem são os mecanismos de classificação mais empregados em pesquisas da literatura cujo intuito é obter SAW de desastres naturais e auxiliar a GD a partir de mensagens compartilhadas em redes sociais (por exemplo, Yin *et al.* (2012), Huang e Xiao (2015), Li *et al.* (2018), Bruijn *et al.* (2020)). Por fim, o modelo de classificação treinado é submetido ao processo de validação cruzada na base de dados de teste, pois necessita-se medir o desempenho do mecanismo computacional quanto as seguintes métricas de avaliação: precisão, *recall*, *f1-score*.

4.2 Planejamento dos Experimentos

Esta etapa da abordagem de Fusão Multimodal possui o intuito de apresentar o planejamento dos experimentos realizados para a criação dos modelos de combinação de dados heterogêneos, no qual são capazes de obter SAW de inundações e auxiliar a GD a partir de mensagens compartilhadas no *Twitter* e informações contextuais. Desse modo, para que seja possível criar essas abordagens de combinação de informações midiáticas, primeiramente necessita-se identificar as possíveis áreas de alagamentos da cidade de São Paulo, portanto o primeiro ex-

perimento denomina-se "**Descoberta das regiões propícias à ocorrência de alagamentos na cidade de São Paulo**". Logo após a identificação das regiões favoráveis ao acontecimento de alagamentos, foca-se no treinamento e avaliação dos modelos de Fusão Multimodal, então o segundo experimento intitula-se "**Treinamento e avaliação dos modelos de Fusão Multimodal**". Por último, há a comparação do desempenho entre os modelos quanto a precisão, *recall* e *f1-score* e a investigação do impacto das informações contextuais no desempenho dos mecanismos de combinação de informações midiáticas, portanto o terceiro experimento realizado nesta abordagem denomina-se "**Comparação entre os modelos de Fusão Multimodal e averiguação do impacto da inclusão de dados contextuais na abordagem**".

Além disso, as configurações do computador responsável pela execução dos três experimentos desta abordagem de obtenção de SAW de desastres naturais a partir de *tweets* e dados contextuais são as seguintes: Sistema Operacional (Ubuntu 19.10); Memória RAM (8GB DDR3 com 1600MHz); CPU (5^a Geração do Intel Core i5-5200U); Placa de Vídeo (NVIDIA(R) GeForce(R) 820M 2GB DDR3).

Dado o exposto, o **primeiro experimento** realizado nesta abordagem concentra-se na identificação das possíveis áreas de enchentes da cidade de São Paulo, além da investigação do impacto no desempenho dos algoritmos de agrupamento quanto a utilização de abordagens empíricas e geo estatísticas para a definição do raio de formação dos *clusters*.

Portanto, para a execução dessas tarefas são aplicados os seguintes algoritmos de *clustering* na base de dados históricos de alagamentos da capital paulista (Tabela 8): DBSCAN; OPTICS; *Hierarchical Agglomerative Clustering* com os critérios de junção de agrupamentos definidos, como: "*Single*", "*Complete*", "*Average*" e "*Ward*". Inclusive, todos esses algoritmos de *clustering* são configurados conforme as diversas distâncias de formação de grupos baseadas em estratégias empíricas e geo estatísticas, visto que o objetivo deste experimento é selecionar o algoritmo de agrupamento e o raio de criação de grupos responsável pela produção das zonas de enchentes mais bem definidas conforme a métrica de avaliação denominada *Silhouette*. Aliás, esta estratégia de seleção de algoritmos de agrupamento e raios de criação de *clusters* mais eficazes para a produção de áreas de alagamentos, foi baseada na abordagem apresentada por Feng e Sester (2018).

Desse modo, uma parcela das distâncias de criação de *clusters* utilizadas neste experimento são baseadas de forma geo estatística no resultado da aplicação do Semivariograma nos dados históricos das ocorrências de alagamentos, já as demais são baseadas empiricamente na variação de um raio mínimo até a dimensão máxima das células de chuva intensa da cidade de São Paulo, ou seja, 100 metros até 5 quilômetros com a frequência aditiva de 100 metros (por exemplo, 100, 200, 300, 400, 500, ..., 5000). Inclusive, o raio de 5 quilômetros é escolhido como limite máximo da abordagem empírica, pois de acordo com Espejo (2016) esse é o tamanho máximo das células de chuva intensa de regiões tropicais. A seguir na Tabela 8, observam-se as características da base de ocorrências históricas de alagamentos da cidade de São Paulo utilizada

neste experimento.

Tabela 8 – dados históricos das ocorrências de alagamentos

coordenadas geográficas das ocorrências históricas de alagamentos	11798
coordenadas geográficas únicas das ocorrências históricas de alagamentos	1433
características	data, período, latitude (*float*), longitude (*float*)
data inicial	01/01/2015
data final	30/10/2018

Fonte: Elaborada pelo autor.

Portanto, nota-se na Tabela 8 que no período de 1 de janeiro de 2015 até 30 de outubro de 2018 houve a incidência de 11798 observações horárias de alagamentos na cidade de São Paulo, além da presença de 1433 coordenadas geográficas únicas de ocorrências históricas de enchentes. Aliás, as características cruciais desta base de dados correspondem a data, período e coordenadas geográficas das incidências dos desastres naturais.

Ademais, este primeiro experimento foi realizado com auxílio da biblioteca de ML escrita na linguagem de programação Python chamada *Scikit-Learn* (versão 0.24.0) (PEDREGOSA *et al.*, 2011), sendo que os parâmetros utilizados pelos algoritmos de agrupamento foram os padrões disponibilizados pela biblioteca, com exceção das distâncias de formação de *clusters* que foram baseadas em abordagens empíricas e geo estatísticas de inferência.

Já o **segundo experimento** realizado nesta abordagem, concentra-se na seleção dos parâmetros responsáveis pela produção dos mecanismos de classificação mais precisos contidos nos modelos de Fusão Multimodal, sendo que esta etapa baseia-se em um conjunto de parâmetros pré-definidos empiricamente a partir da análise das informações contidas na documentação do *Scikit-Learn* (SCIKIT-LEARN, 2021b). Além disso, foca-se no treinamento dos modelos de Fusão Multimodal e seleção das estratégias de conversão de informações simbólicas para numéricas mais eficazes. Ademais, realiza-se o processo de validação cruzada dos diversos modelos de Fusão Multimodal no conjunto de dados de teste. Desse modo, o objetivo deste experimento é descobrir os parâmetros que proporcionam maior precisão aos algoritmos de ML, além de escolher os mecanismos de classificação mais precisos que serão utilizados posteriormente para comparar os modelos de combinação de informações midiáticas e avaliar o impacto da inclusão de dados contextuais na abordagem de Fusão Multimodal.

Dado o exposto, os conjuntos de dados utilizados neste experimento podem ser observados na Tabela 9 e Tabela 10, aliás o processo de criação das bases de dados empregadas para a

seleção dos melhores parâmetros dos algoritmos de ML e treinamento dos modelos de Fusão Multimodal, é detalhado na etapa de elaboração do Conjunto Verdade (Figura 21).

Tabela 9 – Descrição das bases de dados utilizadas para o treinamento dos modelos de Fusão Multimodal

Tipo de Conjunto de Dados	**Características**	**Quantidade de informação**
Fusão Prévia	id (float), data_criacao (date), latitude (float), longitude (float), texto (string), dentro_area_alagamento (int), umidade (float), temperatura (float), precipitacao (float), pressao_atmosferica (float), temperatura_ponto_orvalho (float), rotulo (int)	910
Fusão Tardia - Dados Climáticos	id (float), data_medicao (date), umidade (float), temperatura (float), precipitacao (float), pressao_atmosferica (float), temperatura_ponto_orvalho (float), rotulo (int)	910
Fusão Tardia - Tweets	id (float), data_criacao (string), texto (string), rotulo (int)	910

Fonte: Elaborada pelo autor.

Tabela 10 – Descrição dos dados utilizados no processo de otimização de parâmetros

Algoritmo	**Conjunto de Parâmetros**
RF	bootstrap = [True]; max_depth = [80, 90, 100, 110]; max_features = [2, 3]; min_samples_leaf = [3, 4, 5]; min_samples_split = [8, 10, 12]; n_estimators = [100, 200, 300, 1000].
SVM	C = [0.001, 0.01, 0.1, 1, 10]; gamma = [0.001, 0.01, 0.1, 1, “auto”]; kernel = [“linear”, “rbf”]; decision_function_shape = [“ovo”,“ovr”]; shrinking = [True,False].
NB	var_smoothing = np.logspace(0,-9, num=100).
DT	max_depth = range(1,50); criterion = [“gini”, “entropy”].
RL	penalty = [“l1”,“l2”]; C = [0.001,0.009,0.01,0.09,1,5,10,25]; solver = [“liblinear”,“saga”]; multi_class = [“ovr”,“auto”].

Fonte: Elaborada pelo autor.

Portanto, nota-se na Tabela 9 que as informações textuais e contextuais contidas na base de dados chamada “Fusão Prévia” estão combinadas de forma espaço-temporal, já os dados contidos nos conjuntos de dados denominados “Fusão Tardia - Dados Climáticos” e “Fusão Tardia - Tweet” correspondem respectivamente as informações meteorológicas e aos dados textuais advindos do *Twitter*. Inclusive, as informações detalhadas na Tabela 9 são referentes ao período de 1 de novembro de 2016 até 30 de outubro de 2018 da cidade de São Paulo, aliás utiliza-se nos três conjuntos de dados uma quantia de informações idêntica, pois segundo Faceli *et al.* (2011) em aplicações de ML recomenda-se o balanceamento das bases de dados, pois dessa forma as aplicações não são enviesadas por classes majoritárias de exemplos. Já na

Tabela 10, observa-se o conjunto de dados utilizados no processo de otimização de parâmetros dos algoritmos de ML, inclusive tais informações foram obtidas empiricamente ao analisar a documentação da biblioteca do *Scikit-Learn* (versão 0.24.0) (PEDREGOSA *et al.*, 2011).

Para que seja possível concluir todas as tarefas designadas para este experimento, necessita-se primeiramente pré-processar os dados textuais, assim aplica-se a estratégia de limpeza de texto apresentada na Seção 4.1.5.1 (Algoritmo 2) e posteriormente executa-se a estratégia de transformação de dados apresentada na Seção 4.1.5.1, no qual é responsável por converter os caracteres simbólicos para numéricos com o auxílio das seguintes abordagens: BOW, TF-IDF e *Word Embeddings* do tipo *FastText* (50 e 100 dimensões) e *Word2Vec* (50 e 100 dimensões).

Desse modo, para o treinamento e teste do modelo de "**Fusão Multimodal de modo prévio**" utiliza-se as informações do conjunto de dados contido na Tabela 9 cujo nome é "Fusão Prévia", inclusive o processo de classificação adotado é o definido na Figura 26. Já para o modelo de "**Fusão Multimodal de modo tardio**" existe o treinamento e teste do modelo de identificação de enchentes na base de dados detalhada na Tabela 9 cujo nome é "Fusão Tardia - Dados Climáticos", além disso, há o treinamento e teste do mecanismo computacional capaz de classificar mensagens advindas do *Twitter* na base de dados da Tabela 9 chamada "Fusão Tardia - Tweets". Aliás, ambos processos de classificação adotados seguem o modelo apresentado na Figura 26.

Já no modelo de **Fusão Multimodal de modo híbrido com enfoque na decisão das informações meteorológicas**, existe primeiramente o treinamento e teste de um mecanismo de identificação de enchentes que usufrui das informações presentes na base de dados da Tabela 9 chamada "Fusão Tardia - Dados Climáticos". Logo após, aplica-se o mecanismo de classificação de alagamentos na base de dados da Tabela 9 chamada "Fusão Prévia" e combinam-se os resultados preditos desta base de dados com as respectivas informações textuais e geográficas pré-processadas, assim resultando em um novo conjunto de dados. Por último, treina-se e testa-se um novo mecanismo de classificação neste novo conjunto de dados criado conforme o padrão de elaboração de modelos de classificação apresentado na Figura 26.

Ademais, no modelo de **Fusão Multimodal de modo híbrido com foco na decisão das informações textuais**, há a princípio o treinamento e teste de um mecanismo de classificação de *tweets* que usufrui das informações presentes na base de dados da Tabela 9 chamada "Fusão Tardia - Tweet". Logo após, executa-se o mecanismo de classificação de *tweets* na base de dados da Tabela 9 chamada "Fusão Prévia" e combinam-se os resultados preditos desta base de dados com as respectivas informações meteorológicas e geográficas pré-processadas, assim resultando em uma nova base de dados. Por último, treina-se e testa-se um novo modelo de classificação nesta nova base de dados confeccionada conforme o padrão de elaboração de modelos de classificação apresentado na Figura 26.

Aliás, para que o procedimento de otimização de parâmetros dos algoritmos de ML seja

efetuado, todos os mecanismos de classificação utilizados neste experimento são submetidos a um processo de validação cruzada na base de dados de treinamento (Tabela 9), no qual se variam os parâmetros utilizados pelos algoritmos conforme o conjunto de dados apresentado na Tabela 10. Assim, após a execução de diversas validações cruzadas na base de dados de treinamento com todas as possibilidades de combinação de parâmetros, selecionam-se aqueles que foram responsáveis por proporcionar os modelos de classificação mais precisos (por exemplo, o algoritmo DT com "*max_depth*" equivalente a 30 e "*criterion*" correspondente a "*entropy*").

Além disso, para todos os modelos de classificação empregados nas abordagens de Fusão Multimodal, há a verificação do impacto da utilização de diferentes técnicas de conversão de dados simbólicos para numéricos, quanto aos resultados das seguintes métricas de avaliação empregadas na etapa de teste: precisão, *recall* e *f1-score*. Inclusive, para o treinamento e teste dos mecanismos de classificação de dados foi utilizado a biblioteca de ML escrita na linguagem de programação *Python* chamada *Scikit-Learn* (versão 0.24.0) (PEDREGOSA *et al.*, 2011), aliás o módulo *GridSearchCV*[18] presente nessa biblioteca foi o responsável por auxiliar a etapa de otimização de parâmetros deste experimento.

O **terceiro experimento** realizado nesta abordagem, foca-se na investigação do impacto da inclusão de informações contextuais nos modelos de Fusão Multimodal e na comparação do desempenho dos modelos de combinação de informações midiáticas. Desse modo, primeiramente há o treinamento e teste de um modelo de obtenção de SAW de alagamentos que utiliza somente informações textuais, sendo que as informações textuais utilizadas são as pertencentes a base de dados chamada "Fusão Tardia - Tweet" da Tabela 9, inclusive o processo de confecção do modelo de classificação adotado é o detalhado na Figura 26.

Posteriormente, todos os modelos de Fusão Multimodal e o de classificação unimodal são testados na base de dados de validação presente na Tabela 11, desse modo comparam-se os valores resultantes dos modelos quanto as seguintes métricas de avaliação: precisão, *recall* e *f1-score*. Assim, verifica-se a diferença de desempenho do modelo de Fusão Multimodal que obtém SAW de alagamentos mais precisamente e do modelo de classificação unimodal, pois dessa maneira é possível contabilizar o impacto da inclusão de informações contextuais nas abordagens de Fusão Multimodal. A seguir, na Tabela 11 nota-se a base de dados utilizada neste experimento para a validação dos modelos de treinados.

Por último, observa-se na Tabela 11 a presença de algumas características atípicas, sendo a primeira chamada "**clima_alagamento**", no qual corresponde a verificação da ocorrência de alagamentos na cidade de São Paulo no período que foram efetuadas as medições das variáveis climáticas pelos sensores meteorológicos do INMET. Já a segunda denominada "**tweet_alagamento**", onde se refere ao resultado da classificação manual das mensagens captadas do *Twitter* quanto ao relacionamento com alagamentos (Seção 4.1.3). A terceira característica da base de dados de validação é chamada de "**dentro_area_alagamento**", no qual diz respeito a verificação

[18] url: <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html>

Tabela 11 – Descrição do conjunto de dados de validação

Características	**Quantidade**
identificador (float), texto (string), umidade (float), temperatura (float), precipitacao (float), pressao_atmosferica (float), temperatura_ponto_orvalho (float), clima_alagamento (int), tweet_alagamento (int), dentro_area_alagamento (int), rotulo (int)	224

Fonte: Elaborada pelo autor.

da ocorrência de mensagens publicadas no *Twitter* em áreas propícias ao acontecimento de alagamentos. Já a última característica "**rotulo**", é equivalente ao resultado do processo de combinação de dados midiáticos manual descrito na Subseção 4.1.6.

4.3 Considerações Finais

Este capítulo descreveu as etapas inerentes ao processo de elaboração da abordagem de Fusão Multimodal capaz de obter SAW de desastres naturais e auxiliar a etapa de resposta da GD. Inclusive, neste capítulo houve a descrição das seguintes etapas desta abordagem: Estudo de Caso; Base de Dados; Seleção dos Dados; Classificação Manual das Informações Textuais; Engenharia de Características; Criação do Conjunto Verdade; Modelos de Fusão Multimodal. Assim, para ser possível a criação dos modelos de Fusão Multimodal do tipo prévio, tardio, híbrido com foco na decisão proporcionada pelos dados textuais e híbrido com foco na decisão proporcionada pelas informações climáticas, logo foi apresentado o planejamento dos seguintes experimentos: descoberta das regiões propícias à ocorrência de alagamentos na cidade de São Paulo; treinamento e avaliação dos modelos de Fusão Multimodal; comparação entre os modelos de Fusão Multimodal e averiguação do impacto da inclusão de dados contextuais na abordagem.

CAPÍTULO

5

DESENVOLVIMENTO DA PLATAFORMA DE DETECÇÃO DE POSSÍVEIS VÍTIMAS DE ALAGAMENTOS

Este capítulo possui o objetivo de expor as etapas de desenvolvimento de uma plataforma capaz de obter SAW de desastres naturais e auxiliar a etapa de resposta da GD em tempo real. Desse modo, na Seção 5.1 observa-se a arquitetura do *software*, além disso, há a descrição dos processos inerentes as seguintes camadas: Urls, *Models*, *Views*, *Templates*, Persistência e Detecção de possíveis vítimas de inundações. Por último, são apresentadas as considerações finais deste capítulo (Seção 5.2).

5.1 Abordagem

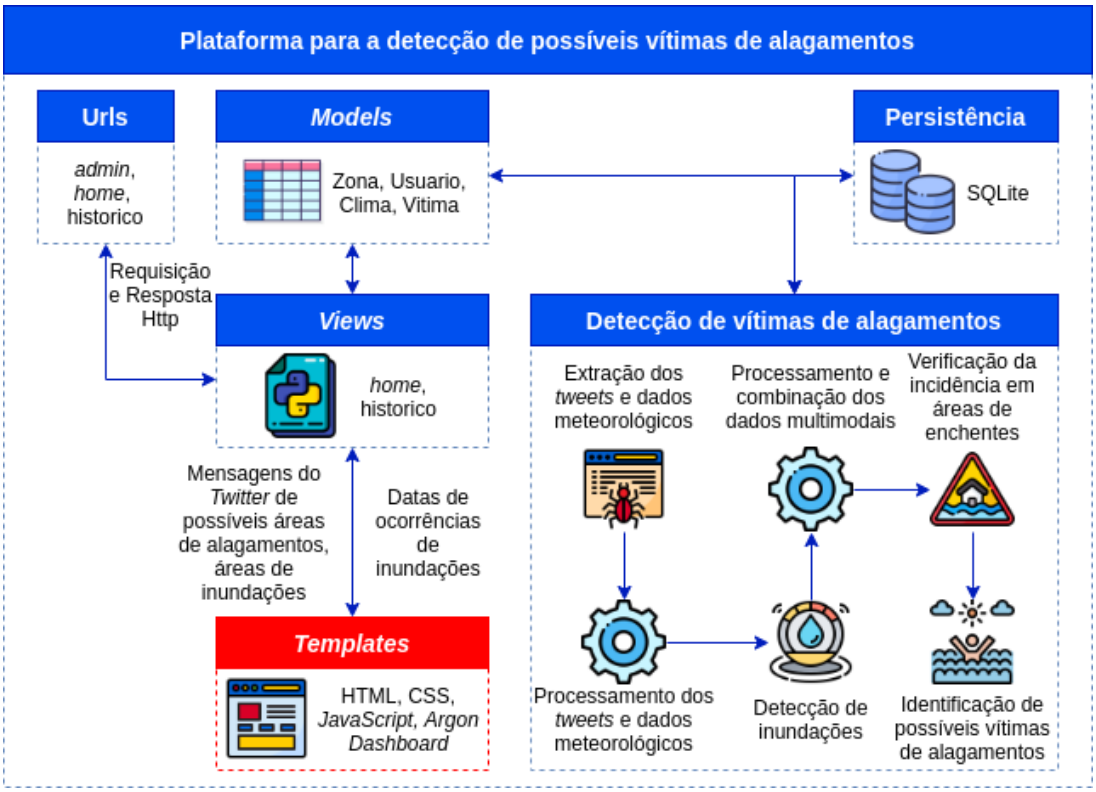
O objetivo desta abordagem é obter SAW de alagamentos, auxiliar a etapa de resposta da GD e consequentemente identificar as possíveis vítimas de inundações da cidade de São Paulo a partir de *tweets*, dados meteorológicos e informações geográficas. Sendo que, este *software* possui diversas funções, como: detecção de possíveis vítimas de enchentes; identificação de áreas de inundações ativas da cidade de São Paulo; disponibilização de uma análise quantitativa das informações processadas pela plataforma de maneira histórica e em tempo real. Desse modo, na Figura 27 é possível observar os usuários deste programa de computador, já na Figura 28 nota-se a arquitetura informacional da plataforma.

Dado o exposto, é possível notar na Figura 27 que os usuários da plataforma de detecção de possíveis vítimas de enchentes da cidade de São Paulo são os Bombeiros, a Defesa Civil e as Organizações não Governamentais (ONGs), inclusive esses potenciais clientes do *software* possuem um papel importante na etapa de resposta da GD, visto que essa etapa é focada na localização e socorro das vítimas de desastres naturais (por exemplo, inundações, terremotos,

Figura 27 – Usuários do *software* e possíveis formas de conexão *Internet*

Fonte: Elaborada pelo autor.

Figura 28 – Arquitetura da plataforma de detecção de possíveis vítimas de alagamentos



Fonte: Elaborada pelo autor.

incêndios, deslizamentos de terra, entre outros). Aliás, esses potenciais usuários podem se conectar com a plataforma através de diferentes formas de conexão com a *Internet* (por exemplo, *Wi-Fi*, 3G, 4G, 5G, entre outras) e acessando um endereço eletrônico específico (por exemplo,

<http://www.sosalagamento.ml/>).

Já na Figura 28, nota-se a arquitetura informacional da plataforma, no qual é constituída de diversas camadas, como: **Urls**, ***Models***, ***Views***, ***Templates***, **Persistência** e **Detecção de vítimas de alagamentos**. Assim, a seguir observam-se as descrições das funcionalidades de cada camada do *software* de identificação de vítimas de inundações da cidade de São Paulo:

- **Urls:** De acordo com Foundation (2021a), a camada de ***Urls*** possui o intuito de proporcionar ao *software* a possibilidade do **redirecionamento** das requisições Http para as respectivas ***Views***, ou seja, há o mapeamento das diversas *Uniform Resource Locators* (URLs) para as funções específicas desta plataforma, no qual estão escritas na linguagem de programação *Python* (FOUNDATION, 2021a).

  Além disso, a conexão com a seção de administração do *software* (***admin***) é realizada somente com as credenciais disponibilizadas pelos administradores da plataforma. Inclusive, na seção de administração é possível a realização do cadastro, atualização e exclusão das informações textuais, meteorológicas, das áreas de inundações da cidade de São Paulo e das possíveis vítimas de alagamentos identificadas.

- ***Models*:** Segundo Foundation (2021c), a camada de ***Models*** contém os campos e os meios de interação com os dados armazenados, ou seja as estratégias para inserir, alterar e excluir as informações do banco de dados. Além disso, cada modelo presente no *software* possui a tarefa de mapear uma tabela do banco de dados (FOUNDATION, 2021c).

- ***Views*:** De acordo com Foundation (2021b), a camada de ***Views*** possui a tarefa de encapsular a lógica responsável pelo processamento das requisições dos usuários e coordenar as respectivas respostas, sendo que as respostas podem ser páginas com o conteúdo escrito em HTML, arquivos do tipo XML, imagens, erros de páginas não encontradas, entre outras. Desse modo, a seguir apresentam-se as funcionalidades das ***Views*** desta plataforma de identificação de possíveis vítimas de alagamentos:

  - ***home*:** A ***View*** chamada ***home*** possui o intuito de retornar para os usuários do *software* as informações necessárias para a obtenção da SAW de inundações da cidade de São Paulo em **tempo real**. Sendo que, para que seja possível essa operação a plataforma capta a data e a hora de acesso do usuário e efetua uma pesquisa em seu banco de dados pelas possíveis vítimas de alagamentos identificadas pela IA em um período de até 4 horas.

    Posteriormente, há o redirecionamento das seguintes informações para a camada de ***Templates***: a quantidade de *tweets* relacionados com inundações compartilhados naquele período; a quantia e as coordenadas das áreas de alagamentos ativas na cidade de São Paulo; a quantidade de possíveis vítimas de alagamentos; as informações públicas sobre as vítimas de enchentes (por exemplo, nome do usuário do

Twitter, data, horário, longitude, latitude e conteúdo textual da mensagem do *Twitter*). Inclusive, os dados públicos são transportados num objeto do tipo *JavaScript Object Notation* (JSON), além disso, há a divulgação dessas informações, pois elas são cruciais para os bombeiros socorrerem as vítimas.

  - **historico:** A ***View*** chamada **historico** possui o objetivo de retornar para os usuários da plataforma as informações **históricas** de possíveis vítimas de enchentes. Dessa forma, necessita-se que o usuário da plataforma comunique o período da publicação das informações desejadas, dessa maneira o *software* realiza uma busca no banco de dados pelas informações referentes a data solicitada.

    Logo após, há o redirecionamento das seguintes informações para a camada de ***Templates***: a quantidade de mensagens do *Twitter* relacionadas com enchentes presentes na data solicitada; a quantia e as coordenadas das áreas de alagamentos ativas no período pesquisado; a quantidade de *tweets* de possíveis vítimas de enchentes na data pesquisada; os dados públicos das vítimas de inundações. Não menos importante, as informações públicas são transmitidas num objeto do tipo JSON, aliás, há a divulgação dessas informações, pois o intuito desta plataforma é auxiliar a etapa de resposta da GD e esses dados são essenciais para os bombeiros ampararem as vítimas.

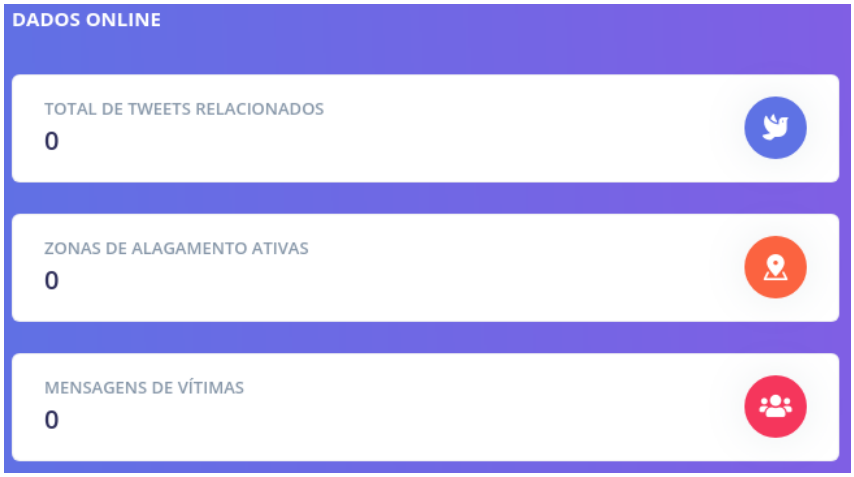
- ***Templates*:** De acordo com Foundation (2021d), a camada de ***Templates*** possui a finalidade de renderizar os dados fornecidos pela camada de ***Views*** em conjunto com as informações advindas de arquivos HTML pré-definidos, assim possibilita-se aos usuários do *software* a visualização dos resultados dos processos de análise de dados de maneira dinâmica. Ademais, os componentes gráficos utilizados nesta camada pertencem ao ***Argon Dashboard***[1], no qual é um *dashboard* de código aberto e gratuito, cuja finalidade é para a utilização em projetos acadêmicos e comerciais.

  A seguir, nas Figura 29 e Figura 30 é possível notar os painéis que proporcionam aos usuários do *software* as **informações em tempo real**, inclusive esses painéis são exibidos a partir do acesso à plataforma de identificação de possíveis vítimas de inundações no dia de **21 de fevereiro de 2021 às 2 horas**. Já nas Figura 31 e Figura 32, nota-se os painéis que proporcionam aos usuários do programa de computados as **informações históricas**, aliás esses painéis são exibidos a partir da pesquisa na plataforma de identificação de possíveis vítimas de enchentes pelas informações captadas no dia **20 de agosto de 2020**. Dessa forma, ambas funcionalidades da plataforma retornam para os usuários *insights* que vão desde a quantia de áreas de alagamentos ativas no período da pesquisa até a localização das vítimas de inundações.

  Além disso, nota-se na Figura 29 que ao acessar o *software* no dia 21 de fevereiro de 2021 não houve ocorrências de *tweets* relacionados com enchentes, não foram detectadas áreas de inundações ativas e o mecanismo de IA da plataforma não identificou potenciais vítimas

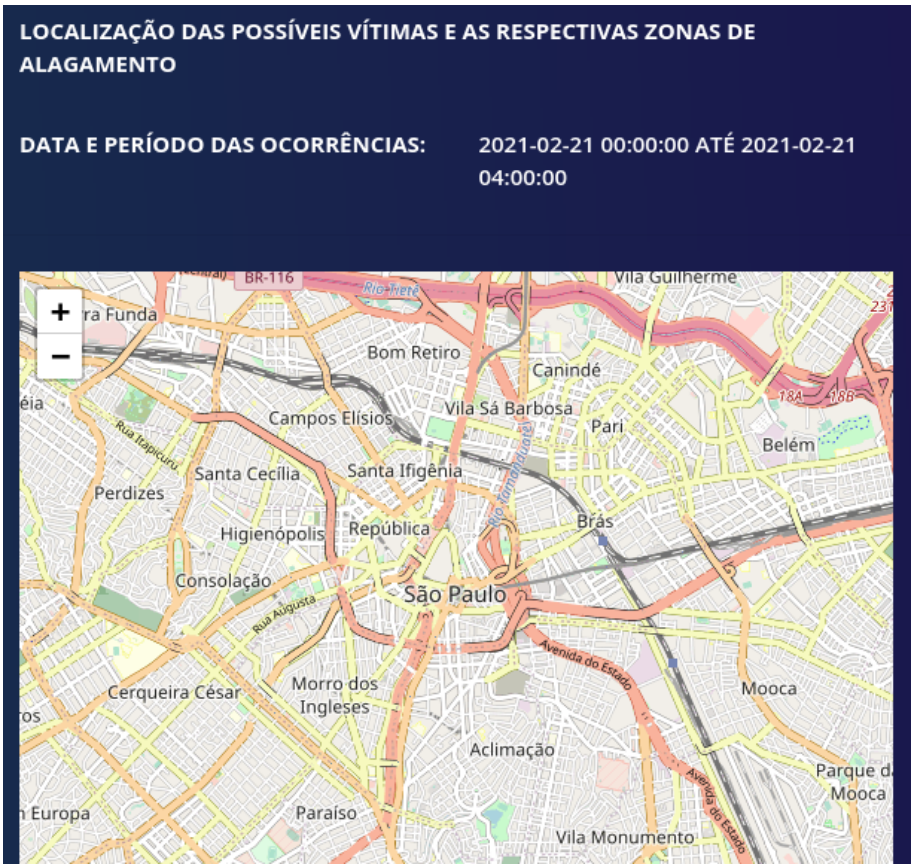
[1] url: <https://www.creative-tim.com/product/argon-dashboard>

Figura 29 – Informações adicionais do serviço de análise de dados online da plataforma de detecção de possíveis vítimas de alagamentos (21/02/2021)



Fonte: Elaborada pelo autor.

Figura 30 – Informações geo localizadas do serviço de análise de dados online da plataforma de detecção de possíveis vítimas de alagamentos (21/02/2021)

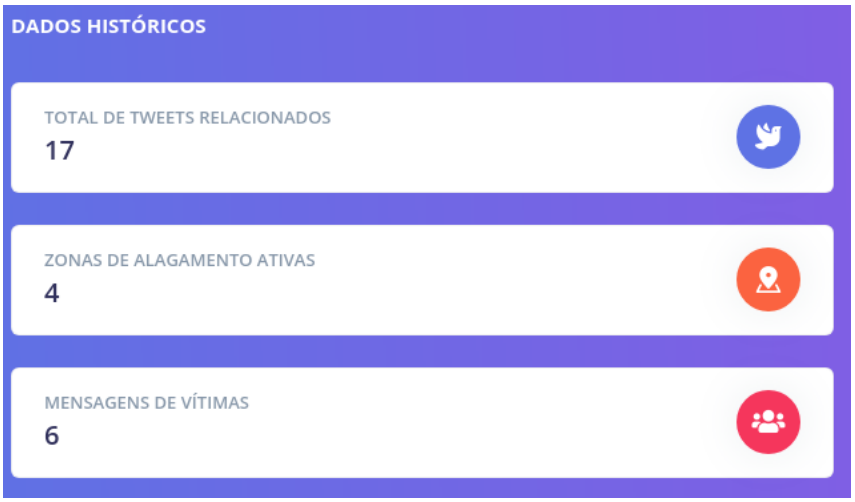


Fonte: Elaborada pelo autor.

de alagamentos. Já na Figura 30, observa-se a presença do mapa da cidade de São Paulo sem VGI, visto que no período da consulta não houve a detecção de possíveis vítimas de

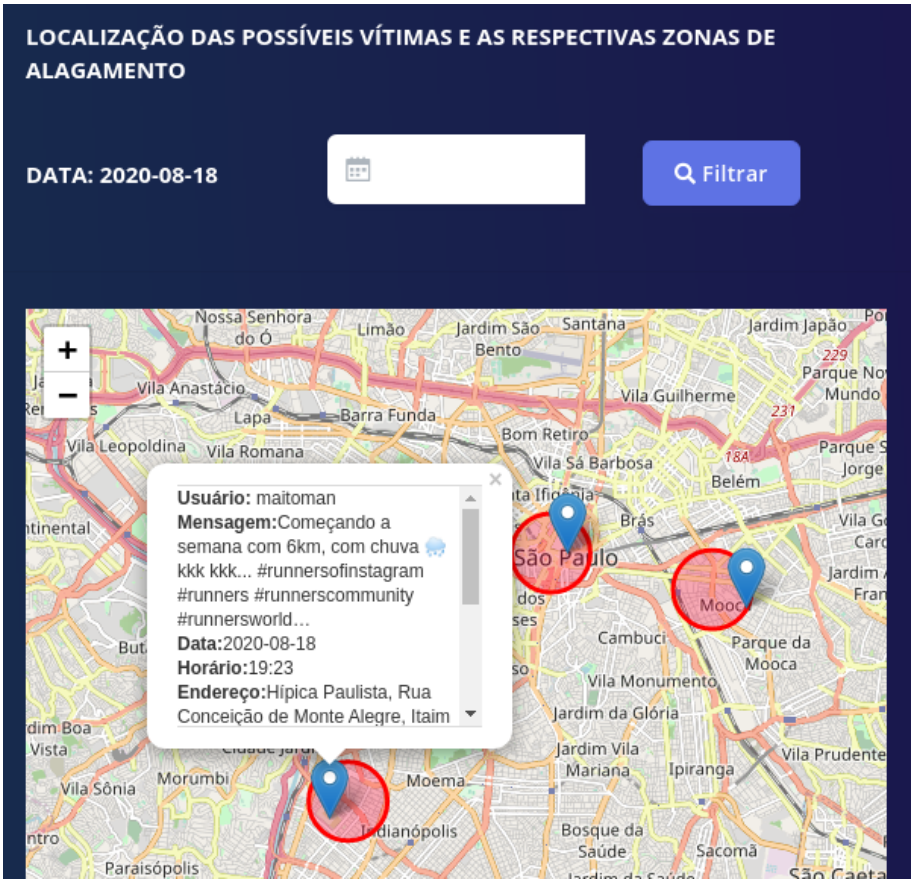
alagamentos pelo modelo de Fusão Multimodal pertencente ao programa de computador.

Figura 31 – Informações adicionais do serviço de análise de dados históricos da plataforma de detecção de possíveis vítimas de alagamentos (20/08/2020)



Fonte: Elaborada pelo autor.

Figura 32 – Informações geo localizadas do serviço de análise de dados históricos da plataforma de detecção de possíveis vítimas de alagamentos (20/08/2020)



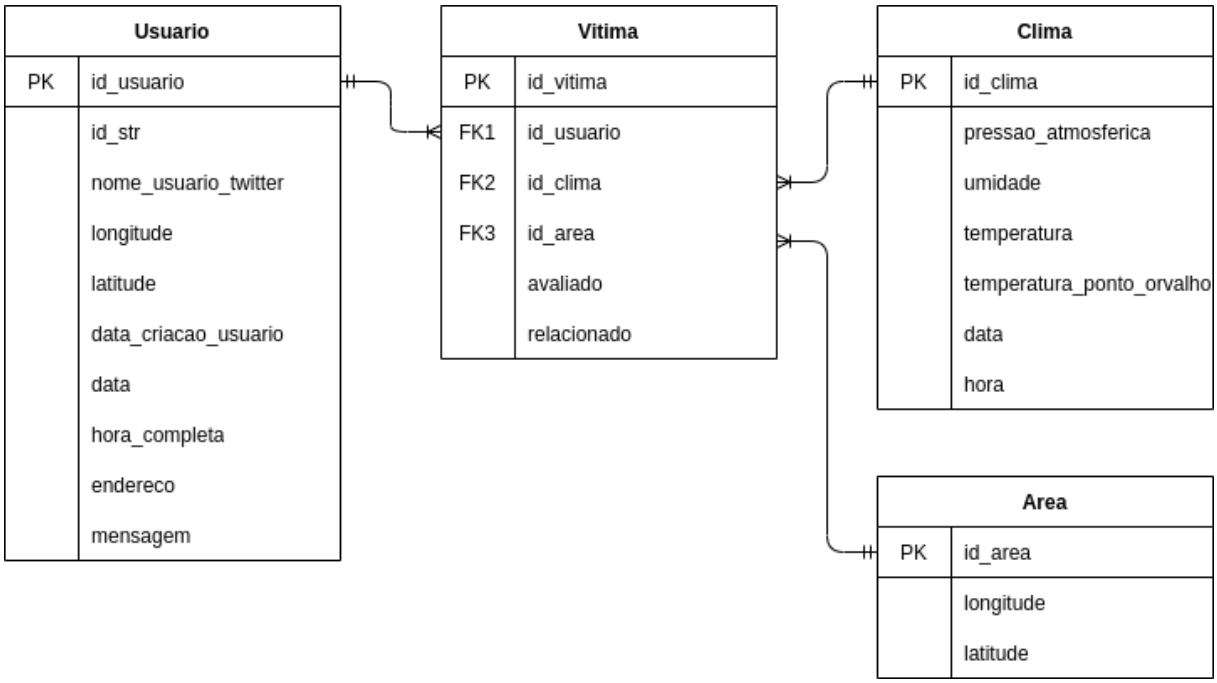
Fonte: Elaborada pelo autor.

Ademais, nota-se na Figura 31 que ao pesquisar as informações históricas do dia 20 de agosto de 2020 obtêm-se **17 *tweets* relacionados com enchentes**. Inclusive, nesta mesma data houve **4 áreas de inundações ativas** e foram detectadas pelo mecanismo de IA **6 mensagens como pertencentes a possíveis vítimas de alagamentos**. Por último, na Figura 32 exibi-se em um mapa as localizações de 3 possíveis vítimas de alagamentos e as respectivas áreas de inundações da cidade de São Paulo, assim ao acessar as informações contidas nos *targets* a plataforma retorna as seguintes informações referentes as potenciais vítimas: **usuário**, **mensagem**, **data**, **horário** e o **endereço**.

- **Persistência:** Esta camada do programa de computador possui o objetivo de persistir os dados previamente mapeados pela camada de ***Models***. Aliás, o banco de dados relacional utilizado na plataforma de detecção de possíveis vítimas de alagamentos foi o SQLite[2].

  Desse modo, em seguida na Figura 33 observa-se o diagrama Entidade Relacionamento (ER) do *software*.

Figura 33 – Diagrama ER da plataforma de detecção de possíveis vítimas de alagamentos a partir de mensagens do *Twitter* e dados contextuais



Fonte: Elaborada pelo autor.

Dado o exposto, é possível notar na Figura 33 que as informações meteorológicas (pressão atmosférica, umidade, temperatura e temperatura do ponto do orvalho) são armazenadas na tabela "**Clima**", inclusive essas informações são captadas constantemente pelo *software* via API do INMET. Já as possíveis áreas de ocorrência de enchentes da cidade de São Paulo são persistidas na tabela "**Area**", aliás o processo de detecção dessas áreas é discutido na

[2] url: <https://www.sqlite.org/>

Subsubseção 4.1.5.2 e os resultados são exibidos na Seção 6.2. Ademais, as informações dos *tweets* (nome de usuário, coordenadas, endereço, data de publicação e mensagem) são armazenadas na tabela "**Usuario**", inclusive esses dados são captados incessantemente via API de *Streamming* do *Twitter* e pré-processados pelo *software*.

Além disso, nota-se na Figura 33 que a tabela "**Vitima**" relaciona-se com as seguintes tabelas: "**Usuario**", "**Clima**" e "**Area**". Sendo que, esses relacionamentos são construídos após a captação das seguintes informações: mensagens do *Twitter* com o conteúdo textual relacionado com situações climáticas (por exemplo, contendo as seguintes palavras-chave: chuva, alagamento, inundação, entre outras); condições meteorológicas do período em que foram publicados os *tweets*; coordenadas geográficas das áreas de alagamentos que as mensagens estão inseridas. Inclusive, inicialmente os atributos "avaliado" e "relacionado" recebem o valor 0, desse modo, **após a execução do mecanismo de IA capaz de identificar possíveis vítimas de inundações**, caso ocorra a **detecção de informações pertencentes a possíveis vítimas de enchentes**, então **atribui-se o valor 1 nos atributos "avaliado" e "relacionado"**, caso contrário atribui-se o valor 1 no atributo "avaliado" e o valor 0 no atributo "relacionado".

Por último, é possível observar na Figura 33 que a cardinalidade entre as tabelas "**Usuario**", "**Clima**" e "**Area**" para a tabela "**Vitima**" é de **1 para muitos**, isto significa que os dados provenientes dessas tabelas podem pertencer a diversas instâncias da tabela **Vitima**, no entanto as informações oriundas da tabela "Vitima" conectam-se somente com instâncias únicas das tabelas "Usuario", "Clima" e "Area".

- **Detecção de vítimas de alagamentos:**

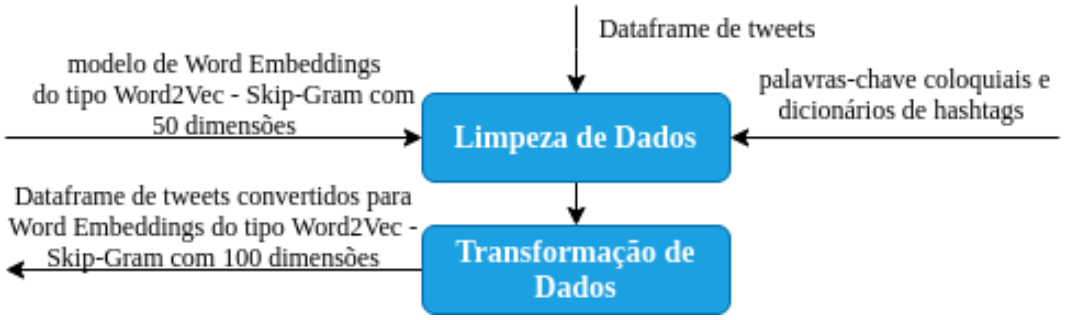
  O processo de detecção de possíveis vítimas de inundações é realizado de forma contínua no *software*, para isso foi desenvolvido um mecanismo computacional capaz de extrair e armazenar os *tweets* da cidade de São Paulo que possuem as palavras-chave relacionadas com situações climáticas, aliás nota-se a metodologia empregada para a seleção das palavras-chave na Subseção 4.1.3 e a respectiva lista na Tabela 5. Inclusive, o *crawler* foi escrito na linguagem de programação *Python* com o auxílio da biblioteca TweetPy[3] e o armazenamento das informações é realizado no banco de dados MongoDB[4].

  Logo após o processo de captação dos *tweets*, há a execução do mecanismo responsável pela identificação de possíveis vítimas de alagamentos na cidade de São Paulo, assim é possível observar na Figura 34 o fluxograma informacional da etapa de pré-processamento textual do mecanismo de identificação.

  De acordo com a Figura 34, as informações textuais são pré-processadas a partir da execução das seguintes etapas: **limpeza de dados** e **transformação de dados**. A etapa de limpeza de dados possui o intuito de remover os ruídos das informações textuais, desse

[3] url: <https://www.tweepy.org/>
[4] url: <https://https://www.mongodb.com/>

Figura 34 – Fluxograma informacional da etapa de pré-processamento do mecanismo de identificação de possíveis vítimas de enchentes



Fonte: Elaborada pelo autor.

modo nota-se na Subsubseção 4.1.5.1 a descrição do processo de confecção do **algoritmo de limpeza de dados** deste trabalho, no qual pode ser observado em Algoritmo 2. Inclusive, este mecanismo usufrui do dicionário de *hashtags* contido no Apêndice A, da lista de palavras-chave coloquiais contidas no Apêndice B e do modelo de *Word Embeddings* do tipo *Word2Vec* treinado pelo NILC, visto que o *corpus textual* analisado contém diversas palavras escritas informalmente que necessitam ser convertidas para a norma culta da Língua Portuguesa.

Ademais, o algoritmo de limpeza de dados realiza as seguintes operações:

- Exclusão de endereços eletrônicos, *stop words*, *emoticons* e caracteres duplicados;
- Tradução de *hashtags*;
- Correção de palavras informais;
- Transformação de elementos textuais para *tokens*.

Assim, após a execução dessas operações o mecanismo de limpeza informacional produz uma coleção de *tweets* sem ruídos e aptos para serem processados pela etapa de transformação de dados.

A etapa de **transformação de dados** possui o objetivo de converter os dados simbólicos para numéricos. Desse modo, as informações resultantes da fase de limpeza de dados são convertidas para *Word Embeddings* do tipo *Word2Vec* - Skip-Gram com 100 dimensões. Assim, o mecanismo de transformação de dados produz uma coleção de *tweets* transformados para vetores densos numéricos, aptos para serem combinados com as demais informações heterogêneas e processados pelo modelo de ML capaz de detectar de vítimas de enchentes.

Em seguida, há um exemplo da execução do algoritmo de pré-processamento textual, no qual na Tabela 12 há um conjunto de *tweets* contendo palavras relacionadas com situações climáticas, já na Tabela 13 existem os resultados da execução do algoritmo de limpeza

de dados e na Tabela 14 há os resultados da fase de conversão de dados simbólicos para numéricos.

Tabela 12 – Conjunto de *tweets* utilizado como exemplo

identificador	tweet
1	#avenidapaulista #saopaulo #saopaulocity #sp #splovers #amorpaulista #brasil #cidadedagaroa...
2	#boanoite que #chuva boa e as pessoas reclamando... Pode chover todos os dias... Trabalhar no...

Fonte: Elaborada pelo autor.

Tabela 13 – Conjunto de *tweets* resultantes da etapa de limpeza de dados

identificador	tweet
1	[avenida paulista, são paulo, são paulo, são paulo, amando em são paulo, brasil, cidade da garoa]
2	[boa noite, chuva, boa, pessoas, reclamando, pode, chover, todos, dias, trabalhar]

Fonte: Elaborada pelo autor.

Tabela 14 – Exemplo da conversão do conjunto de *tweets* para *Word Embeddings* - *Word2Vec* do tipo Skip-Gram com 100 dimensões

identificador	tweet
1	[0.2558218538761139, ... ,0.2905715107917785]
2	[-0.0275687109678983, ... ,0.2420501410961151]

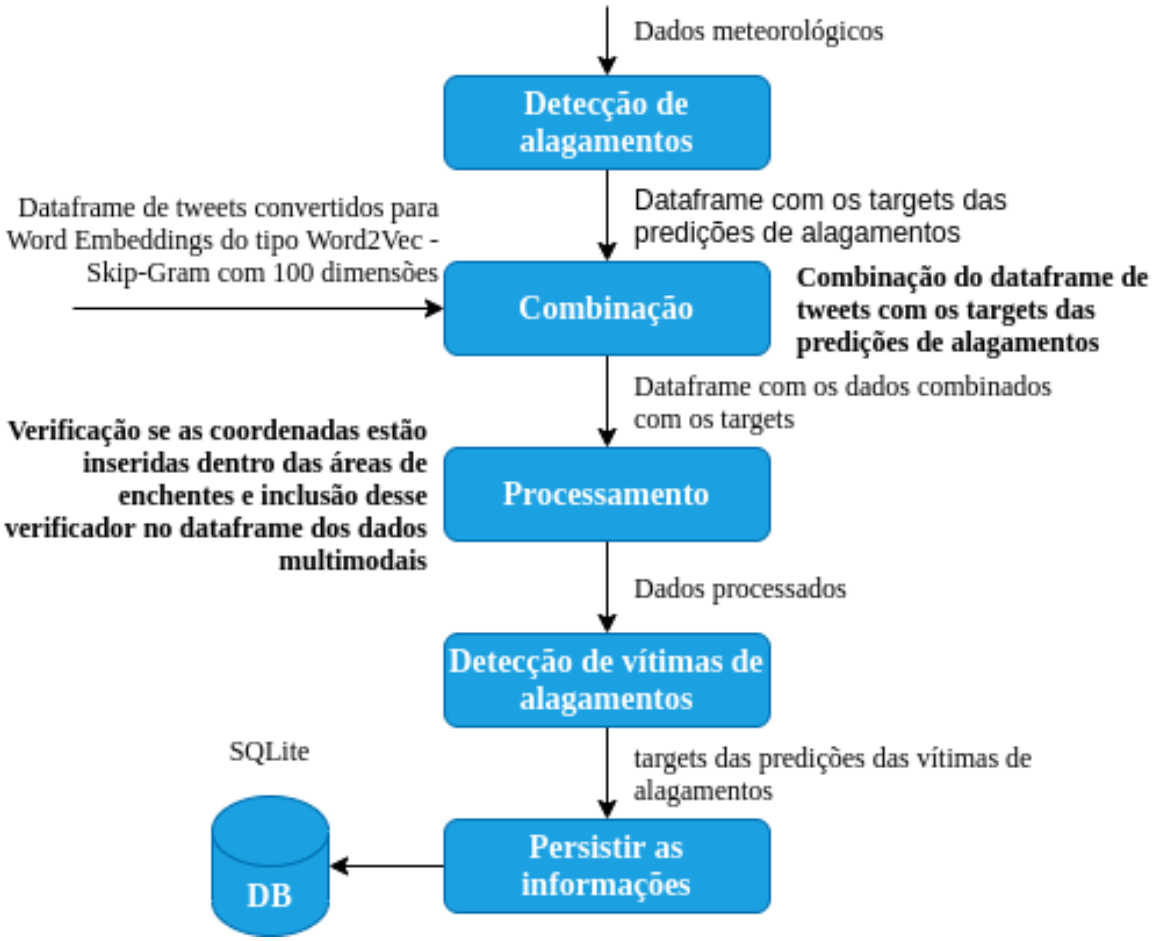
Fonte: Elaborada pelo autor.

A seguir na Figura 35 nota-se as etapas posteriores do mecanismo de identificação de possíveis vítimas de alagamentos, assim há as fases de **detecção de alagamentos**, **combinação de informações heterogêneas**, **identificação de vítimas de enchentes** e **armazenamento dos dados**.

Segundo a Figura 35, é possível notar que as informações numéricas resultantes da etapa de conversão de dados são **combinadas de maneira linear com os resultados da fase de detecção de inundações**, ou seja, combinam-se os vetores densos numéricos produzidos anteriormente com os *targets* das predições de alagamentos do período em que foram publicadas as mensagens do *Twitter*.

Inclusive, o mecanismo computacional capaz de **identificar enchentes** usufrui de dados climáticos (**temperatura**, **temperatura do ponto do orvalho**, **umidade** e **pressão atmosférica**), possui **80,03% de precisão** e foi treinado com as informações climáticas captadas do INMET do período de 1 de novembro de 2016 até 30 de outubro de 2018 da cidade de São Paulo,

Figura 35 – Fluxograma informacional das etapas de combinação, processamento e identificação de possíveis vítimas de alagamentos do *software*



Fonte: Elaborada pelo autor.

além disso, na Seção 4.2 detalham-se os experimentos realizados para a construção do modelo de ML capaz de detectar a presença de enchentes, já na Seção 6.3 apresentam-se os resultados obtidos.

Logo após, conforme se nota na Figura 35 há a **verificação se as informações multimodais combinadas pertencem às áreas de alagamentos da cidade São Paulo e a inclusão desse verificador no *dataframe* das informações heterogêneas**, ou seja, caso as coordenadas geográficas das mensagens do *Twitter* estejam inseridas em áreas propícias ao acontecimento de alagamentos, então adiciona-se o valor 1 na coluna "ìnside_cluster" da linha analisada do *dataframe* dos dados multimodais, caso contrário há a inclusão de 0.

Aliás, essas regiões propícias ao acontecimento de inundações foram descobertas a partir do agrupamento dos pontos de alagamentos notificados pelo CGE de janeiro de 2015 até outubro de 2018, no qual o algoritmo de *clustering* aplicado é o *Agglomerative Hierarchical Clustering* do tipo *Ward* com 900 metros de alcance, ademais, na Seção 4.2 detalham-se os experimentos realizados para a identificação das regiões da cidade de São Paulo propícias à ocorrência de

inundações, já na Seção 6.2 apresentam-se os resultados obtidos.

Posteriormente, as informações multimodais combinadas (dados textuais pré-processados, *targets* de predição de alagamentos e verificador de disseminação de mensagens em áreas de enchentes da cidade de São Paulo) são submetidas ao processo de **detecção pelo modelo de ML capaz de identificar possíveis vítimas de inundações**, assim caso detectado uma vítima de alagamento armazena-se na instância analisada da tabela "Vitima" os atributos "avaliado" e "relacionado" com o valor 1, caso contrário persisti-se o atributo "avaliado" com o valor 1 e o "relacionado" com o valor 0.

Além disso, o **mecanismo de IA** empregado para **detectar possíveis vítimas de inundações** possui **84,70% de precisão**, foi treinado com o algoritmo de ML chamado DT e os dados utilizados para o treinamento são de um *ground truth* criado neste trabalho, no qual é constituído de *tweets* pré-processados, *targets* resultantes da predição de alagamentos e dos verificadores de incidência de mensagens em áreas propícias ao acontecimento de enchentes (Subseção 4.1.6), ademais, na Seção 4.2 detalham-se os experimentos realizados para a elaboração do modelo de Fusão Multimodal, já na Seção 6.3 e Seção 6.4 apresentam-se os resultados obtidos.

Conclui-se que esta camada do *software* é responsável pela captação e processamento das informações multimodais, além da detecção de alagamentos e de possíveis vítimas de enchentes. Sendo que, todos os mecanismos computacionais foram escritos na linguagem de programação *Python* e utilizaram desde bibliotecas para o auxílio na captação de informações advindas de redes socias (por exemplo, *TweetPy*[5]) e PLN (por exemplo, NLTK[6]) até o auxílio no treinamento e teste dos modelos de ML (por exemplo, *Scikit-Learn*[7]).

5.2 Considerações Finais

Este capítulo detalhou o desenvolvimento da plataforma de detecção de possíveis vítimas de alagamentos da cidade de São Paulo, além de atender um dos objetivos específicos desta dissertação que é a elaboração de um *software* capaz de identificar possíveis vítimas de enchentes em tempo real a partir de um modelo de Fusão Multimodal que utiliza informações textuais e contextuais. Aliás, a plataforma possui diversas funções, como: identificação de possíveis vítimas de enchentes; detecção de regiões de alagamentos ativas da capital paulista; disponibilização de uma análise quantitativa dos dados processados pelo *software* de maneira histórica e em tempo real. Por último, esta plataforma possui o intuito de auxiliar a etapa de resposta da GD e fornecer SAW de alagamentos da cidade de São Paulo em tempo real para que os bombeiros possam socorrer possíveis vítimas.

[5] url: <https://www.tweepy.org/>
[6] url: <https://www.nltk.org/>
[7] url: <https://scikit-learn.org/stable/>

CAPÍTULO

6

RESULTADOS E DISCUSSÕES

Este capítulo possui o objetivo de apresentar as métricas de avaliação empregadas nos experimentos desta dissertação (Seção 6.1), além de exibir os seguintes resultados: **Descoberta das regiões propícias ao acontecimento de alagamentos** (Seção 6.2), desse modo nesta seção existe a exibição dos resultados inerentes aos processos de escolha da distância mais adequada para a elaboração das áreas de alagamentos da cidade de São Paulo e análise do desempenho dos algoritmos de *clustering* quanto a aplicação nas ocorrências históricas de enchentes; **Treinamento e avaliação dos modelos de Fusão Multimodal** (Seção 6.3), assim nesta seção apresentam-se os resultados dos processos de treinamento e teste dos modelos de Fusão Multimodal, além da avaliação do impacto das estratégias de transformação de dados textuais para numéricos quanto ao desempenho dos mecanismos de classificação; **Comparação entre os modelos de Fusão Multimodal e averiguação do impacto da inclusão de dados contextuais na abordagem** (Seção 6.4), portanto nesta seção observa-se o resultado da inclusão de informações meteorológicas e geográficas nos modelos de combinação de informações midiáticas, aliás notam-se também os resultados da comparação entre os modelos de Fusão Multimodal quanto as seguintes métricas de avaliação: precisão, *recall* e *f1-score*. Por último, são apresentadas as considerações finais deste capítulo (Seção 6.5).

6.1 Métricas de Avaliação

As métricas de avaliação utilizadas nos experimentos são baseadas em pesquisas encontradas na literatura que possuem o objetivo de elaborar mecanismos computacionais capazes de: classificar mensagens advindas de redes sociais (por exemplo, Rosa *et al.* (2011), Ashktorab *et al.* (2014), Salas, Georgakis e Petalas (2017)), agrupar informações geográficas voluntárias (por exemplo, Sparks *et al.* (2020)) e combinar informações heterogêneas (por exemplo, Poria *et al.* (2016), Pouyanfar *et al.* (2019), Bruijn *et al.* (2020)).

Dessa forma, as métricas de avaliação utilizadas são as seguintes: *Silhouette*, precisão,

recall e *f1-score*. A *Silhouette* foi empregada para analisar o desempenho dos mecanismos computacionais responsáveis pela identificação das possíveis áreas de alagamentos da cidade de São Paulo, já as demais foram utilizadas para analisar o desempenho das abordagens de Fusão Multimodal e dos mecanismos de classificação de informações heterogêneas.

A métrica de avaliação chamada *Silhouette* possui o objetivo de determinar o quão bem formados estão os *clusters* identificados pelos algoritmos de *clustering*, ou seja, analisa-se a distância entre os elementos do espaço amostral e os agrupamentos que eles não estão inseridos (SCIKIT-LEARN, 2021a). Dessa forma, os agrupamentos que são considerados como bem definidos possuem a característica de alcançar resultados próximos de 1 (um positivo) ao serem submetidos ao processo de avaliação pela métrica *Silhouette*, já os *clusters* que são interpretados como mal definidos alcançam resultados de *Silhouette* próximos de -1 (um negativo) (SPARKS *et al.*, 2020). Não menos importante, os resultados de *Silhouette* próximos de 0 indicam a existência de sobreposição de *agrupamentos* (SCIKIT-LEARN, 2021a).

Além disso, as métricas de avaliação empregadas no processo de análise do desempenho das estratégias de Fusão Multimodal e dos algoritmos de classificação, possuem a característica de serem comumente aplicadas em problemas de duas classes (por exemplo, a classificação de *tweets* em relacionados a alagamentos ou não, a identificação dos dados meteorológicos que indicam a presença de enchentes ou não, entre outros). Portanto, segundo Faceli *et al.* (2011) nesta categoria de desafio uma classe é rotulada como positiva (+), já a outra como negativa (-). Aliás, na Tabela 15 observa-se a matriz de confusão que descreve as possíveis soluções para esta categoria de desafio, inclusive, as classes previstas pelos mecanismos de classificação se encontram nas colunas e as verdadeiras classes das amostras estão contidas nas linhas.

Tabela 15 – Matriz de confusão, no qual descreve os desafios presentes nesta dissertação.

	+	-
+	VP	FN
-	FP	VN

Fonte: Faceli *et al.* (2011).

Dado o exposto, observa-se na Tabela 15 que o elemento denominado VP (Verdadeiro Positivo) é equivalente à quantidade de exemplos classificados como positivo corretamente pelos mecanismos de classificação, já o VN (Verdadeiro Negativo) corresponde a quantia de exemplos classificados como negativo de forma correta (FACELI *et al.*, 2011). Ademais, nota-se também a presença dos elementos FN (Falso Negativo) e FP (Falso Positivo) na Tabela 15, portanto o item FN corresponde a quantidade de exemplos classificados erroneamente como negativo, pois a classe verdadeira dos exemplos do espaço amostral é positiva, já o item FP equivale à quantidade de exemplos identificados erroneamente como positivo, visto que a classe verdadeira das amostras utilizadas pelo mecanismo de classificação é negativa (FACELI *et al.*, 2011).

Desse modo, a partir das informações contidas na matriz de confusão apresentada na

Tabela 15, deriva-se as métricas de avaliação utilizadas nesta dissertação para computar o desempenho dos mecanismos de classificação e dos modelos de Fusão Multimodal. Portanto, segundo Faceli *et al.* (2011) a precisão é equivalente à quantidade de amostras rotuladas verdadeiramente como positivas pelos mecanismos computacionais em relação a quantia de elementos classificados como positivos (Equação 6.1). Já a métrica de avaliação chamada *recall*, de acordo com Faceli *et al.* (2011) é equivalente à quantidade de acertos que os mecanismos computacionais obtiveram na classe positiva dos exemplos do espaço amostral (Equação 6.2). Por fim, segundo Salas, Georgakis e Petalas (2017) a métrica de avaliação denominada *f1-score* corresponde a média harmônica entre os resultados da precisão e o *recall* (Equação 6.3).

$$\text{Precisão} = \frac{TP}{TP+FP} \tag{6.1}$$

$$Recall = \frac{TP}{TP+FN} \tag{6.2}$$

$$F1\text{-}score = \frac{2 \times P \times R}{P+R} \tag{6.3}$$

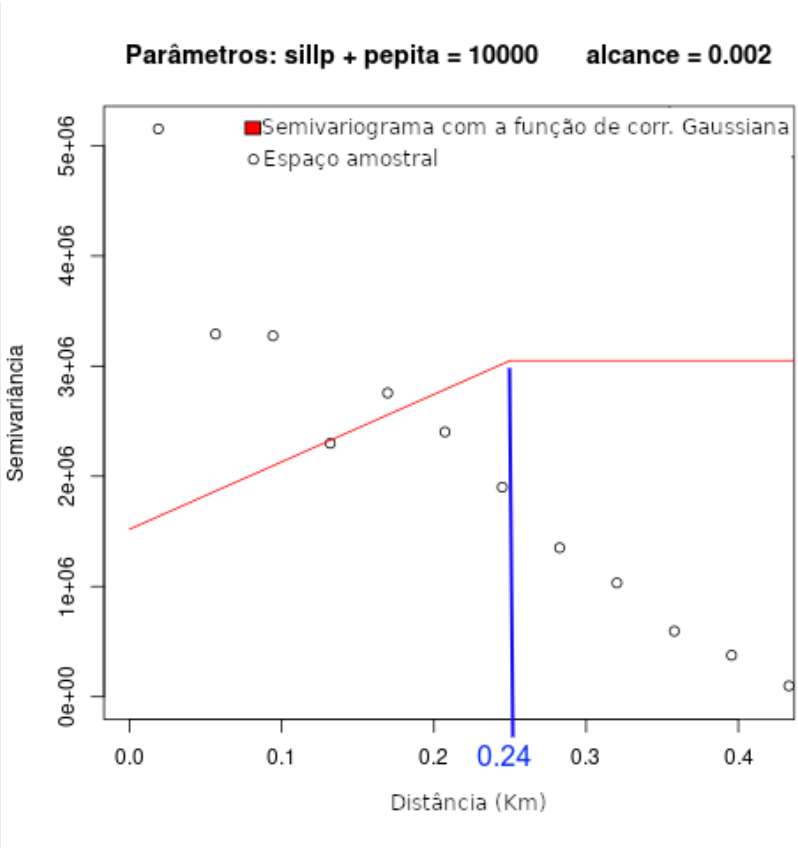
6.2 Descoberta das regiões propícias à ocorrência de alagamentos na cidade de São Paulo

O processo de planejamento da identificação de possíveis áreas de inundações da cidade de São Paulo pode ser observado na Seção 4.2, desse modo este experimento é extremamente importante, pois para a elaboração dos modelos de Fusão Multimodal necessita-se da combinação de dados textuais com contextuais, portanto é necessário identificar os *tweets* que estão localizados geograficamente em regiões propícias ao acontecimento de alagamentos, visto que com base na hipótese de Tobler (1970), as mensagens compartilhadas no *Twitter* que são relacionadas com inundações tendem estar contidas geograficamente em áreas que frequentemente ocorrem enchentes.

Para que seja possível a execução do agrupamento das informações históricas de alagamentos, primeiramente identificamos as dimensões elegíveis das áreas de alagamentos da cidade de São Paulo, sendo que uma parcela dessas dimensões são baseadas em uma estratégia empírica, já a outra é baseada em uma estratégia geo estatística. A estratégia empírica utilizada neste experimento foi baseada na abordagem de identificação de áreas de inundações proposta por Feng e Sester (2018), onde os autores derivaram diversos raios a partir da variação de um raio mínimo pré-definido até a dimensão máxima das células de chuva intensa da região analisada. Dessa forma, como nesta dissertação o estudo de caso é a cidade de São Paulo, então foram utilizadas distâncias entre 100 metros até 5 quilômetros. Já a estratégia geo estatística empregada neste experimento, baseia-se na aplicação da técnica chamada Semivariograma nas ocorrências

históricas de alagamentos, inclusive o intuito desta técnica geo estatística é obter a distância de estabilização dos dados geográficos (ISAAKS; SRIVASTAVA, 1989). Portanto, na Figura 36 nota-se o resultado da aplicação do Semivariograma nos dados históricos de inundações da cidade de São Paulo do período de 1 de janeiro de 2015 até 30 de outubro de 2018 (Tabela 8).

Figura 36 – Resultado da aplicação do Semivariograma nas informações geográficas históricas de inundações da cidade de São Paulo



Fonte: Elaborada pelo autor.

Dado o exposto, observa-se na Figura 36 que o "*range*" obtido na aplicação do Semivariograma nos dados históricos de inundações corresponde a 240 metros, ou seja, a distância de estabilização dos dados geográficos do espaço amostral. Desse modo, foram executados os seguintes algoritmos de *clustering* nos dados históricos de alagamentos detalhados na Tabela 8: DBSCAN, OPTICS, *Agglomerative Clustering* com o critério de ligação do tipo Single, Complete, *Average* e *Ward*. Inclusive, todos os mecanismos de agrupamento foram configurados conforme as possíveis distâncias de identificação de áreas de enchentes apresentadas na Tabela 16.

Neste experimento foram realizados 306 testes de identificação de áreas suscetíveis a alagamentos da cidade de São Paulo, assim na Tabela 17 é possível observar os resultados em ordem decrescente de *Silhouette*, inclusive nesta tabela constam somente os resultados mais relevantes de cada algoritmo de *clustering* e de cada categoria de distância de formação de

Tabela 16 – Distâncias elegíveis para a criação das áreas de alagamentos da cidade de São Paulo

Distâncias obtidas empiricamente (metros)	**Distância obtida geo estatisticamente (metros)**
100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3700, 3800, 3900, 4000, 4100, 4200, 4300, 4400, 4500, 4600, 4700, 4800, 4900, 5000	240

Fonte: Elaborada pelo autor.

agrupamentos. Além disso, também é possível notar na Tabela 17 as seguintes informações sobre os testes deste experimento: qual algoritmo de agrupamento utilizado para identificar as áreas suscetíveis a ocorrência de enchentes; a quantidade de áreas de alagamentos que foram identificadas pelos algoritmos não supervisionados; a categoria de distância de formação de grupos utilizada nos testes (por exemplo, empírica ou geo estatística); as distâncias máximas de formação de *clusters* empregadas pelos mecanismos de agrupamento; o quão bem formadas estão as regiões propícias ao acontecimento de alagamentos, ou seja, o resultado da aplicação da métrica de avaliação chamada *Silhouette* nas regiões de inundações identificadas.

Tabela 17 – Resultados dos processos de identificação de alagamentos

Ranking	**Algoritmo**	**Silhouette**	**Quantidade de áreas de alagamentos identificadas**	**Distância elegível para a criação dos grupos (metros)**	**Categoria da distância de criação de grupos**
1	Agglomerative Clustering (Average)	0,4788	271	900	empírico
2	Agglomerative Clustering (Complete)	0,4765	425	700	empírico
8	Agglomerative Clustering (Single)	0,4692	5	3500	empírico
9	DBSCAN	0,4692	5	3500	empírico
51	Agglomerative Clustering (Average)	0,4045	730	240	estatístico
67	Agglomerative Clustering (Ward)	0,3839	873	5000	empírico
177	OPTICS	0,0931	103	3100	empírico

Fonte: Elaborada pelo autor.

Observa-se na Tabela 17 que o teste de identificação de possíveis regiões suscetíveis a ocorrência de alagamentos que alcançou o melhor desempenho de *Silhouette* e usufruiu de uma distância máxima de formação de grupos obtida **empiricamente**, foi o executado com o algoritmo não supervisionado chamado ***Agglomerative Clustering***, sendo o critério de conexão entre grupos o ***Average***, aliás esta abordagem identificou 271 áreas suscetíveis ao acontecimento de enchentes na cidade de São Paulo, além de adotar 900 metros como dimensão máxima das áreas de alagamentos e obter **0,4788** de ***Silhouette***.

Em contrapartida, também é possível notar na Tabela 17 que o teste de identificação de regiões suscetíveis ao acontecimento de inundações que obteve o melhor desempenho de *Silhouette* e usufruiu de uma distância máxima de formação de grupos obtida **estatisticamente**, foi o executado pelo algoritmo de agrupamento chamado ***Agglomerative Clustering***, sendo o critério de ligação de grupos o ***Average***, inclusive esta abordagem detectou 730 possíveis regiões de enchentes na cidade de São Paulo e adotou 240 metros como dimensão máxima das áreas de alagamentos, além disso, esta abordagem de agrupamento obteve **0,4045** de ***Silhouette***.

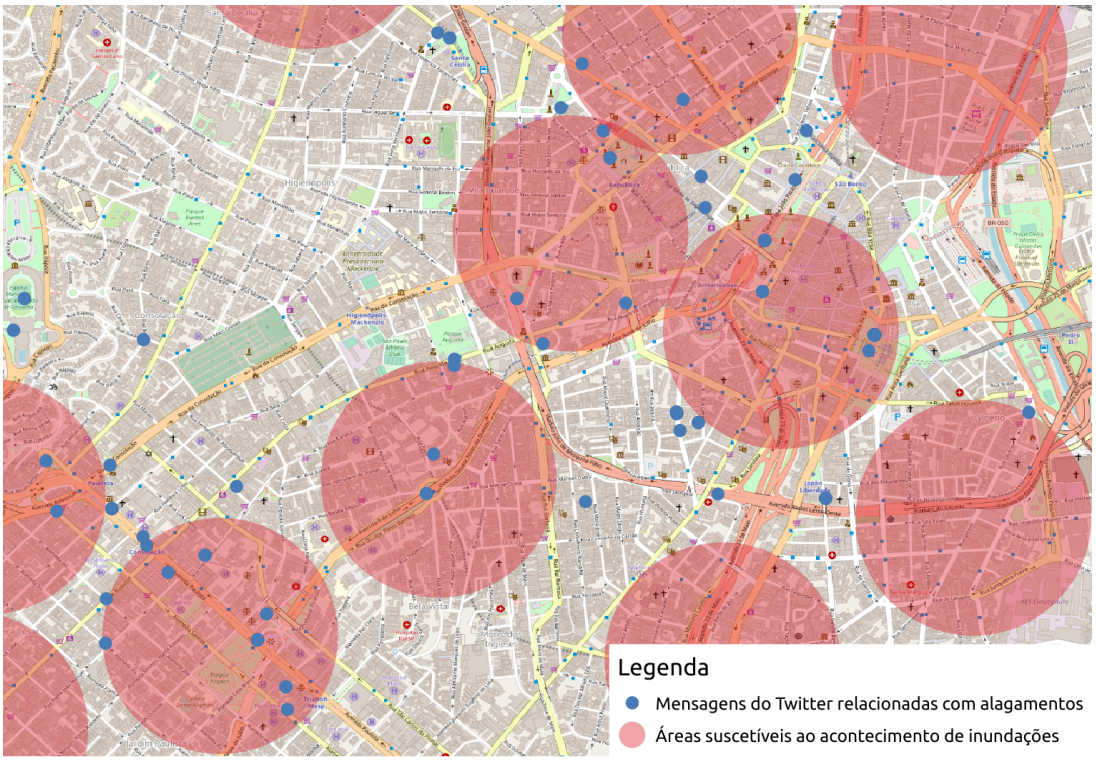
A abordagem de identificação de possíveis áreas de alagamentos da cidade de São Paulo que é a primeira colocada da Tabela 17, detectou grupos mais bem definidos, pois o valor de *Silhouette* alcançado por essa abordagem de agrupamento é mais próximo de 1 do que as demais abordagens de *clustering* testadas neste experimento. Além disso, também é possível observar na Tabela 17 que grande parte dos mecanismos de *clustering* que usufruíram de distâncias de formação de *clusters* obtidas empiricamente alcançaram desempenhos superiores em comparação com os mecanismos que utilizaram distâncias de criação de *clusters* obtidas de maneira estatística, pois de acordo Yin e Li (2001) as enchentes são fenômenos naturais ocasionados pela intervenção inadequada dos seres humanos na natureza, além de que o Semivariograma é uma técnica geo estatística que necessita que as informações geográficas dos desastres naturais possuam dependência espacial para que elas possam ser analisadas (CIGAGNA *et al.*, 2015; METHERON, 1963). Portanto, por esse motivo que os testes que utilizaram distâncias obtidas de forma estatística possuem o desempenho de *Silhouette* inferior quando comparados com as abordagens que usufruíram de distâncias captadas empiricamente, porque não existe dependência espacial entre as ocorrências históricas de alagamentos captadas do CGE-SP.

Ademais, somente no mês de fevereiro de 2017 aconteceram 194 ocorrências de enchentes na cidade de São Paulo, inclusive neste mesmo período houve o compartilhamento de 1224 mensagens no *Twitter* contendo palavras-chave relacionadas com fenômenos naturais (Tabela 5), sendo que após o processo de classificação manual dessas mensagens foram encontrados 875 *tweets* relacionados com alagamentos e 349 mensagens alusivas às inundações (Subseção 4.1.4). Aliás, as áreas da cidade de São Paulo que se destacam pela quantidade de incidências de fenômenos naturais estão localizadas na região central da metrópole (por exemplo, Sé, República, São Bento, Santa Efigênia, Anhangabaú, entre outras).

Dessa forma, é possível observar nas Figura 37 e Figura 38 as regiões suscetíveis ao

acontecimento de alagamentos na cidade de São Paulo e as mensagens publicadas no *Twitter* que são alusivas às inundações do mês de fevereiro de 2017 da capital paulista. Além disso, as áreas de alagamentos presentes na Figura 37 foram identificadas pelo algoritmo de *clustering* chamado *Agglomerative Clustering*, no qual foi configurado com o critério de ligação dos *clusters* do tipo *Average* e a distância máxima de formação de agrupamentos definida em 900 metros. Inclusive, visualizam-se os resultados produzidos por esse mecanismo de agrupamento em uma parcela do mapa da cidade de São Paulo, pois esse algoritmo de *clustering* apresenta o desempenho de *Silhouette* mais expressivo de todos os testes realizados neste experimento, além de ser um exemplo satisfatório da utilização de uma distância de criação de grupos obtida empiricamente.

Figura 37 – Áreas suscetíveis ao acontecimento de alagamentos com o raio de 900 metros e *tweets* geo localizados relacionados com inundações

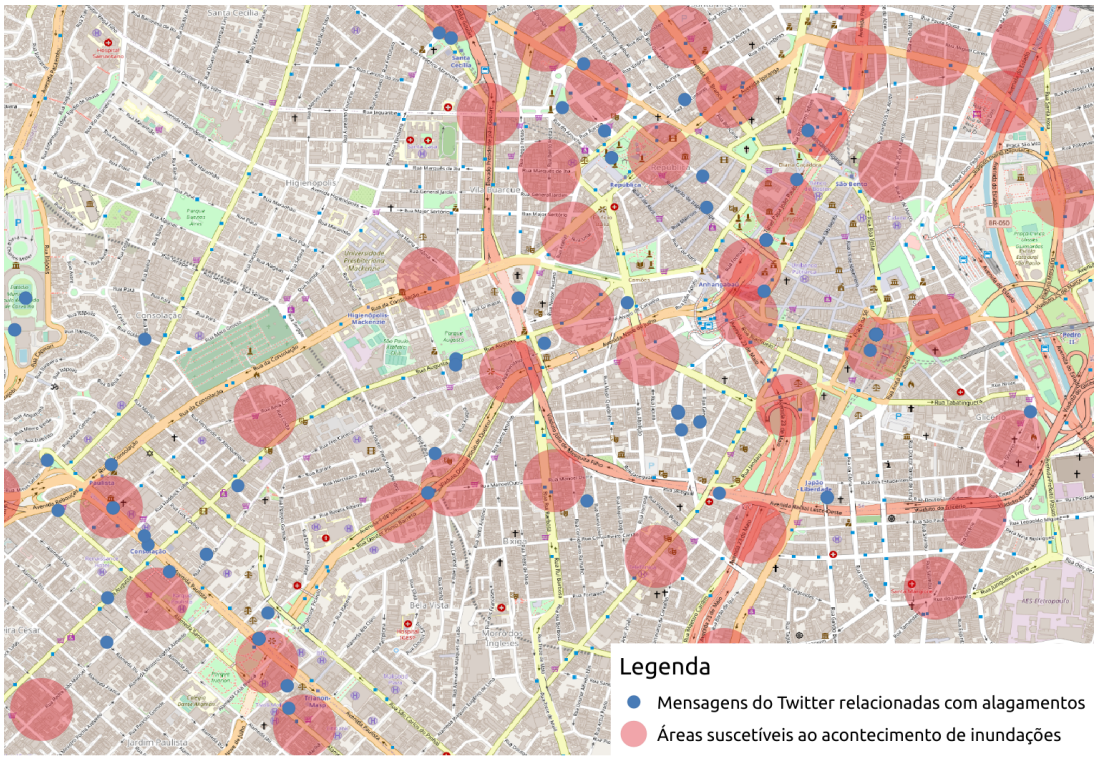


Fonte: Elaborada pelo autor.

Ademais, é possível observar na Figura 38 as possíveis áreas de enchentes identificadas pelo algoritmo de *clustering* chamado *Agglomerative Clustering*, onde foi configurado com o critério de ligação de grupos do tipo *Complete* e a distância máxima de formação de agrupamentos definida em 240 metros. Inclusive, foi escolhido os resultados desse mecanismo de agrupamento para serem exibidos em uma parte do mapa da cidade de São Paulo, pois esse algoritmo de *clustering* é responsável pelo desempenho de *Silhouette* mais satisfatório entre os testes realizados com distâncias de formação de *clusters* obtidas de forma geo estatística.

Dado o exposto, observa-se que as regiões suscetíveis a ocorrência de alagamentos presentes na Figura 37 são mais bem definidas que as da Figura 38, pois o desempenho de *Silhou-*

Figura 38 – Áreas suscetíveis ao acontecimento de enchentes com o raio de 240 metros e *tweets* geo localizados relacionados com alagamentos



Fonte: Elaborada pelo autor.

ette do mecanismo de agrupamento responsável pela identificação dos grupos com a dimensão de até 900 metros é superior ao algoritmo de *clustering* responsável pelo reconhecimento de agrupamentos com o raio de até 240 metros, além de que na Figura 38 existe uma quantidade exacerbada de sobreposições entre as áreas de alagamentos e um número excessivo de *clusters*. Já na Figura 37, nota-se que há uma quantia razoável de sobreposições e o número de áreas de enchentes produzidas na cidade de São Paulo é inferior em comparação com a quantidade de agrupamentos com a dimensão de até 240 metros (Figura 38). Ademais, observa-se que as mensagens publicadas no *Twitter* que são relacionadas com inundações possuem a tendência de se localizarem geograficamente nas regiões propícias ao acontecimento de enchentes apresentadas na Figura 37 em detrimento dos agrupamentos da Figura 38.

Por último, como elucidado anteriormente, este experimento é extremamente importante, pois ele proporciona a identificação de áreas de alagamentos da cidade de São Paulo que são imprescindíveis para a execução do processo de combinação de dados multimodais, inclusive **adota-se nesta dissertação as regiões favoráveis ao acontecimento de alagamentos identificadas pelo *Agglomerative Clustering* com o critério de conexão de grupos do tipo *Average* e a dimensão máxima de criação de *clusters* definida em 900 metros**, pois esse mecanismo de agrupamento obteve o desempenho de *Silhouette* mais significativo entre todos os experimentos realizados, além das áreas de enchentes identificadas serem as mais bem definidas.

6.3 Treinamento e avaliação dos modelos de Fusão Multimodal

Os processos de planejamento do treinamento e teste dos mecanismos de combinação de informações midiáticas, seleção dos parâmetros responsáveis pelos melhores desempenhos dos mecanismos de classificação e seleção das abordagens de conversão de informações textuais para numéricas mais efetivas, podem ser observados na Seção 4.2, desse modo esses procedimentos são importantes, pois são essenciais para a elaboração dos modelos de Fusão Multimodal, no qual são capazes de obter SAW de inundações e auxiliar a etapa de resposta da GD.

Primeiramente, para que seja possível o treinamento e teste das abordagens de combinação de informações midiáticas, logo necessita-se a execução da etapa de processamento das informações textuais, sendo que essa etapa seguiu a metodologia elucidada na Subsubseção 4.1.5.1, inclusive esse método é focado na utilização de estratégias de limpeza de dados e transformação de dados. Além disso, esta etapa deste experimento possui o intuito de identificar o impacto da utilização de diferentes técnicas de conversão de informações simbólicas para numéricas (por exemplo, BOW, TF-IDF e *Word Embeddings* do tipo *Fast Text* e *Word2Vec* com 50 e 100 dimensões) quanto as seguintes métricas de avaliação: precisão, *recall*, *f1-score*.

Posteriormente, existe a elaboração dos modelos de Fusão Multimodal, sendo que esta fase deste experimento seguiu a metodologia elucidada na Seção 4.2. Desse modo, para o desenvolvimento da abordagem de combinação de informações multimodais de modo prévio, a princípio são fundidas as informações textuais, meteorológicas e geográficas ao nível de vetor de recursos, logo após aplica-se o processo de classificação genérico exibido na Figura 26 (ver Subseção 4.1.7). Já para a elaboração da abordagem de fusão de dados multimodais de modo tardio, aplica-se o processo de classificação genérico exibido na Figura 26 individualmente nos dados textuais e climáticos, em seguida combinam-se os resultados das camadas de decisão com as informações geográficas (ver Subseção 4.1.7). Aliás, para o desenvolvimento do modelo de Fusão Multimodal de modo híbrido com foco na decisão captada com as informações textuais, necessita-se fundir os dados textuais ao nível de decisão com os demais dados ao nível de vetor de recursos, logo após aplica-se o processo de classificação genérico exibido na Figura 26 (ver Subseção 4.1.7). Por último, para a elaboração do modelo de Fusão Multimodal de modo híbrido com foco na decisão obtida com os dados meteorológicos, necessita-se combinar as informações meteorológicas ao nível de decisão com os demais dados ao nível de recursos, em seguida executa-se o processo de classificação genérico apresentado na Figura 26 (ver Subseção 4.1.7).

Desse modo, os resultados do processo de otimização de parâmetros dos algoritmos de classificação do modelo de **Fusão Multimodal de modo prévio** podem ser observados na Tabela 18. Assim, na Tabela 18 nota-se que o impacto da execução do processo de otimização de parâmetros do SVM é de 44,43, logo os parâmetros encontrados pelo *GridSearchCV* destacam-se pela aderência as informações multimodais analisadas. No entanto, também é possível notar

na Tabela 18 que o efeito do aperfeiçoamento dos parâmetros do RF é de -0,45, pois os parâmetros definidos como padrão pelo *Scikit-Learn* para esse algoritmo são mais favoráveis aos dados heterogêneos analisados do que os encontrados pelo módulo *GridSearchCV* da biblioteca *Scikit-Learn*. Portanto, devido a este teste, utiliza-se na etapa de treinamento do modelo de Fusão Multimodal de modo prévio os seguintes algoritmos de classificação com os parâmetros aperfeiçoados: SVM, DT, NB, LR. Já quanto ao mecanismo de classificação chamado RF, emprega-se na etapa de treinamento os padrões obtidos com a biblioteca *Scikit-Learn*, porque o seu desempenho foi inferior com os parâmetros otimizados.

Tabela 18 – Resultados da otimização dos parâmetros utilizados pelos algoritmos de classificação utilizados no modelo de Fusão Multimodal de modo prévio

Algoritmo	**Melhores parâmetros**	**Precisão com os parâmetros padrões**	**Precisão com os parâmetros otimizados**	**Impacto da otimização dos parâmetros**
SVM	‘C’: 10, ‘decision_function_shape’: ‘ovo’, ‘gamma’: 1, ‘kernel’: ‘rbf’, ‘shrinking’: True	0,2282	0,6725	44,43
RF	‘bootstrap’: True, ‘max_depth’: 100, ‘max_features’: 3, ‘min_samples_leaf’: 3, ‘min_samples_split’: 10, ‘n_estimators’: 200	0,8547	0,8502	-0,45
DT	’criterion’: ‘gini’, ‘max_depth’: 3	0,7628	0,8554	9,26
LR	‘C’: 25, ‘multi_class’: ‘ovr’, ‘penalty’: ‘l1’, ‘solver’: ’liblinear’	0,6314	0,8338	20,24
NB	‘var_smoothing’: 0.0002848	0,8351	0,8396	0,45

Fonte: Elaborada pelo autor.

Assim, os resultados mais significativos do processo de treinamento dos mecanismos de classificação do modelo de **Fusão Multimodal de modo prévio** são exibidos na Tabela 19. Inclusive, nesta fase deste experimento foram executados 50 testes de *ten-fold cross validation*, onde são as aplicações dos diversos mecanismos de classificação detalhados na Seção 4.2 para cada categoria de transformação de dados textuais para numéricos (por exemplo, BOW, TF-IDF, *Word Embeddings* do tipo *FastText* e *Word2Vec*) no espaço amostral multimodal de treinamento. Aliás, os resultados na íntegra dos testes desta etapa deste experimento podem ser observados no Apêndice E.

Dado o exposto, nota-se na Tabela 19 que o algoritmo de classificação do modelo de combinação de informações midiáticas de modo prévio que obteve o melhor resultado na etapa de treinamento é o **RF** com **0,8715** de precisão e utilizando **BOW** como abordagem de transformação de dados simbólicos para numéricos. Aliás, desde a fase de limpeza de dados,

Tabela 19 – Melhores resultados do processo de treinamento do modelo de Fusão Multimodal de modo prévio

Classificação	Algoritmo	Estratégia de transformação de dados textuais para numéricos	Precisão	Recall	F1-Score
1	RF	BOW	0,8715	0,8667	0,8691
2	DT	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,8702	0,8444	0,8550
3	NB	BOW	0,8659	0,8659	0,8659
4	DT	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8635	0,8278	0,8465
5	RF	TF-IDF	0,8626	0,8547	0,8562

Fonte: Elaborada pelo autor.

transformação de informações textuais para numéricas, combinação de dados multimodais e treinamento dos algoritmos de classificação, o tempo gasto de processamento pela máquina utilizada para a execução dos testes desta etapa deste experimento foram de cerca de **1 hora e 9 minutos**.

Já os resultados do processo de otimização de parâmetros dos mecanismos de classificação do modelo de **Fusão Multimodal de modo tardio** são exibidos nas Tabela 20 e Tabela 21. Desse modo, na Tabela 20 existem os resultados do aperfeiçoamento dos parâmetros da abordagem de identificação de *tweets* relacionados com alagamentos, assim observa-se que os piores desempenhos de otimização de parâmetros são pertencentes aos algoritmos de classificação chamados RF e DT, visto que o RF apresenta um resultado de -2,73 de precisão ao ser submetido ao processo de otimização e o DT um resultado de -0,31 de precisão ao ser executado o procedimento de aperfeiçoamento, portanto se nota que os parâmetros escolhidos pelo módulo *GridSearchCV* da biblioteca *Scikit-Learn* não são aderentes ao espaço amostral de *tweets* correlatos com inundações. Já os demais algoritmos da Tabela 20, apresentam resultados semelhantes ao serem submetidos ao processo de otimização de parâmetros, visto que não há diferença entre os resultados obtidos com a utilização de parâmetros captados de forma padrão da biblioteca *Scikit-Learn* e ao utilizar o módulo *GridSearchCV*. Portanto, devido a este teste, utiliza-se na etapa de treinamento da abordagem de identificação de *tweets* relacionados com alagamentos os seguintes mecanismos de classificação com os parâmetros aperfeiçoados: SVM, LR, NB. Já quanto os algoritmos de classificação denominados RF e DT, utilizam-se os parâmetros padrões obtidos com a biblioteca *Scikit-Learn*, porque os seus desempenhos foram inferiores com os parâmetros aprimorados.

Além disso, na Tabela 21 notam-se os resultados do processo de aperfeiçoamento dos parâmetros dos mecanismos de classificação da abordagem de identificação de alagamentos a

Tabela 20 – Resultados da otimização dos parâmetros utilizados pelos algoritmos de classificação empregados no modelo de identificação de mensagens do *Twitter* relacionadas com alagamentos da abordagem de Fusão Multimodal de modo tardio

Algoritmo	**Melhores parâmetros**	**Precisão com os parâmetros padrões**	**Precisão com os melhores parâmetros**	**Impacto da otimização dos parâmetros**
SVM	'C': 1, 'decision_function_shape': 'ovo', 'gamma': 1, 'kernel': 'rbf', 'shrinking': True	0,8035	0,8035	0,00
RF	'bootstrap': True, 'max_depth': 90, 'max_features': 3, 'min_samples_leaf': 4, 'min_samples_split': 8, 'n_estimators': 200	0,7889	0,7616	-2,73
DT	'criterion': 'entropy', 'max_depth': 4	0,6552	0,6521	-0,31
LR	'C': 1, 'multi_class': 'ovr', 'penalty': 'l2', 'solver': 'saga'	0,7821	0,7821	0,00
NB	'var_smoothing': 0.0012328	0,7357	0,7357	0,00

Fonte: Elaborada pelo autor.

partir de dados meteorológicos. Dessa forma, observa-se na Tabela 21 que o impacto da otimização de parâmetros com o resultado mais significativo advém do algoritmo de classificação chamado SVM, visto que após o aperfeiçoamento dos parâmetros houve um acréscimo de 24,29 na precisão do modelo de identificação de inundações, logo conclui-se que os parâmetros encontrados pelo módulo *GridSearchCV* da biblioteca *Scikit-Learn* são extremamentos aderentes ao espaço amostral das informações meteorológicas. Em contrapartida, também é possível observar na Tabela 21 que o impacto da otimização de parâmetros que apresenta o pior desempenho entre os testes realizados advém do algoritmo chamado RF, visto que houve um decréscimo de -1,69 na precisão ao utilizar os parâmetros obtidos com o *GridSearchCV* em comparação com os fornecidos de forma padrão pela biblioteca *Scikit-Learn*, assim nota-se que os parâmetros otimizados para este algoritmo não são aderentes as informações climáticas presentes na base de dados de treinamento. Dessa maneira, devido a este teste, emprega-se na etapa de treinamento do mecanismo de identificação de enchentes os seguintes algoritmos de classificação com os parâmetros aperfeiçoados: SVM e LR. Já quanto os algoritmos de classificação chamados RF, DT e NB, utiliza-se na etapa de treinamento os padrões obtidos com a biblioteca *Scikit-Learn*, porque seus desempenhos foram inferiores com os parâmetros otimizados.

Assim, os resultados mais significativos dos processos de treinamento das abordagens que fazem parte do modelo de **Fusão Multimodal de modo tardio** são exibidos nas Tabela 22 e Tabela 23. Sendo que, nesta etapa deste experimento para a elaboração do mecanismo de **identi-**

Tabela 21 – Resultados da otimização dos parâmetros utilizados pelos algoritmos de classificação empregados no modelo de identificação de alagamentos da abordagem de Fusão Multimodal de modo tardio

Algoritmo	**Melhores parâmetros**	**Precisão com os parâmetros padrões**	**Precisão com os parâmetros otimizados**	**Impacto da otimização dos parâmetros**
SVM	'C': 10, 'decision_function_shape': 'ovo', 'gamma': 1, 'kernel': 'rbf', 'shrinking': True	0,4298	0,6727	24,29
RF	'bootstrap': True, 'max_depth': 80, 'max_features': 3, 'min_samples_leaf': 3, 'min_samples_split': 8, 'n_estimators': 100	0,8034	0,7865	-1,69
DT	'criterion': 'gini', 'max_depth': 27	0,7832	0,7794	-0,38
LR	'C': 5, 'multi_class': 'ovr', 'penalty': 'l1', 'solver': 'liblinear'	0,6961	0,7211	2,50
NB	'var_smoothing': 0.0043287	0,7324	0,7220	-1,03

Fonte: Elaborada pelo autor.

ficação de mensagens do *Twitter* relacionadas com enchentes foram necessários a execução de 50 testes de *ten-fold cross validation*, no qual são as aplicações de diversos algoritmos de ML para cada categoria de transformação de dados simbólicos para numéricos (por exemplo, BOW, TF-IDF, *Word Embeddings* do tipo *FastText* e *Word2Vec*) na base de dados de treinamento, denominada como "Fusão Tardia - Tweets" (Tabela 9). Dessa forma, conforme a Tabela 22 o algoritmo que obteve o melhor desempenho na tarefa de identificar *tweets* relacionados com alagamentos da etapa de treinamento é o **SVM** com **0,7950** de precisão e a estratégia de transformação de dados empregada é baseada em ***Word Embeddings* do tipo *Fast Text* da categoria *Skip-Gram* com 100 dimensões** do NILC (HARTMANN *et al.*, 2017). Inclusive, o resultado de todos os testes realizados nesta etapa deste experimento podem ser notados no Apêndice F.

Além disso, é possível observar na Tabela 23 os resultados mais expressivos do processo de treinamento dos algoritmos de classificação da abordagem de **identificação de alagamentos a partir de informações meteorológicas**. Inclusive, para que seja possível a conclusão desta etapa foi executado o *ten-fold cross validation* com os algoritmos de ML definidos na Seção 4.2 na base de dados de treinamento, no qual é denominada como "Fusão Tardia - Dados Climáticos" (Tabela 9). Dessa maneira, de acordo com a Tabela 23 o algoritmo de classificação que obteve o melhor desempenho é o **RF** com **0,8003** de precisão. Não menos importante, o tempo gasto pela máquina responsável pela execução dos testes dos mecanismos de identificação de mensagens do *Twitter* relacionadas com enchentes e identificação de alagamentos a partir de dados

Tabela 22 – Melhores resultados do processo de treinamento do mecanismos de identificação de mensagens do *Twitter* relacionadas com alagamentos da abordagem de Fusão Multimodal de modo tardio

Classificação	**Algoritmo**	**Estratégia de transformação de dados textuais para numéricos**	**Precisão**	**Recall**	**F1-Score**
1	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,7950	0,7874	0,7912
2	SVM	BOW	0,7915	0,7885	0,7887
3	SVM	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,7873	0,7715	0,7786
4	SVM	TF-IDF	0,7855	0,7816	0,7836
5	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,7839	0,7701	0,7769

Fonte: Elaborada pelo autor.

meteorológicos é de cerca de **1 hora e 38 minutos**.

Tabela 23 – Melhores resultados do processo de treinamento do mecanismo de identificação de alagamentos da abordagem de Fusão Multimodal de modo tardio

Classificação	**Algoritmo**	**Precisão**	**Recall**	**F1-Score**
1	RF	0,8003	0,7912	0,7957
2	DT	0,7346	0,7308	0,7327
3	LR	0,7283	0,7253	0,7268
4	NB	0,7038	0,6923	0,6968
5	SVM	0,6721	0,6154	0,6174

Fonte: Elaborada pelo autor.

Já o modelo de **Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelos dados textuais**, possui o intuito de combinar as informações contextuais da base de dados de treinamento com as informações resultantes do modelo de identificação de *tweets* relacionados com enchentes, aliás instrui-se esse mecanismo de classificação de mensagens do *Twitter* na etapa de treinamento da abordagem de combinação de dados multimodais de modo tardio. Inclusive, exibi-se de maneira resumida os resultados do processo de treinamento do modelo de identificação de *tweets* relacionados com inundações na Tabela 22, além disso, é possível observar os resultados dessa abordagem na íntegra no Apêndice F.

Desse modo, os resultados do processo de otimização de parâmetros dos mecanismos de classificação do modelo de combinação de dados multimodais de modo híbrido com foco na decisão conferida pelos dados textuais são exibidos na Tabela 24. Sendo que, nota-se na Tabela 24 que o melhor desempenho de otimização de parâmetros desta abordagem de combinação

multimodal é alcançado pelo SVM, visto que o impacto da otimização dos parâmetros é de 44,01, logo conclui-se que as características encontradas pelo *GridSearchCV* são aderentes ao espaço amostral multimodal. Em contrapartida, o resultado menos significativo encontrado foi o exibido pelo NB, pois o impacto do ajuste dos parâmetros é de -0,96, portanto se nota que os parâmetros ajustados para este modelo não são aderentes ao espaço amostral heterogêneo. Dessa maneira, devido a este teste, emprega-se na etapa de treinamento do modelo de Fusão Multimodal de modo híbrido com foco na decisão promovida pelas mensagens do *Twitter* os seguintes algoritmos de classificação com os parâmetros aperfeiçoados: SVM, RF, DT e LR. Já quanto os mecanismos de classificação denominados NB, utilizam-se os parâmetros padrões obtidos com a biblioteca *Scikit-Learn*, pois os seus desempenhos foram inferiores após o ajuste.

Tabela 24 – Resultados da otimização dos parâmetros utilizados pelos algoritmos de classificação empregados no modelo de Fusão Multimodal de modo híbrido com foco na decisão viabilizada pelos dados textuais

Algoritmo	**Melhores parâmetros**	**Precisão com os parâmetros padrões**	**Precisão com os parâmetros otimizados**	**Impacto da otimização dos parâmetros**
SVM	‘C’: 10, ‘decision_function_shape’: ‘ovo’, ‘gamma’: 1, ‘kernel’: ‘rbf’, ‘shrinking’: True	0,2077	0,6478	44,01
RF	‘bootstrap’: True, ‘max_depth’: 110, ‘max_features’: 2, ‘min_samples_leaf’: 4, ‘min_samples_split’: 12, ‘n_estimators’: 100	0,8170	0,8446	2,76
DT	‘criterion’: ‘entropy’, ‘max_depth’: 5	0,8058	0,8406	3,48
LR	‘C’: 25, ‘multi_class’: ‘ovr’, ‘penalty’: ‘l1’, ‘solver’: ‘liblinear’	0,6336	0,8390	20,53
NB	‘var_smoothing’: 0,0000811	0,8586	0,8490	-0,96

Fonte: Elaborada pelo autor.

Assim, para a elaboração do mecanismo de **combinação de dados multimodais de modo híbrido com foco na decisão das informações textuais** realiza-se nesta etapa o *ten-fold cross validation*, no qual são aplicações de diversos algoritmos de ML no conjunto de dados denominado “Fusão Prévia” (Tabela 9). Aliás, utiliza-se nesta etapa deste experimento a estratégia de transformação de dados responsável pelo algoritmo de classificação com melhor desempenho do mecanismo de identificação de *tweets* relacionados com alagamentos. Dessa maneira, a seguir na Tabela 25 observam-se os resultados obtidos com a fase de treinamento do modelo de Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelos *tweets*.

Dado o exposto, é possível notar na Tabela 25 que o mecanismo de classificação res-

Tabela 25 – Melhores resultados do processo de treinamento do modelo de Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelos dados textuais

Classificação	**Algoritmo**	**Estratégia de transformação de dados textuais para numéricos**	**Precisão**	**Recall**	**F1-Score**
1	DT	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8578	0,8333	0,8417
2	NB	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8473	0,8222	0,8342
3	RF	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8394	0,8389	0,8392
4	LR	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8183	0,8056	0,8118
5	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,6673	0,6145	0,6414

Fonte: Elaborada pelo autor.

ponsável pelo melhor desempenho foi o **DT** com **0,8578** de precisão e utilizando a estratégia de transformação de dados baseada em ***Word Embeddings*** **do tipo** ***FastText*** **da categoria** ***Skip-Gram*** **com 100 dimensões** do NILC (HARTMANN *et al.*, 2017). Aliás, desde a execução do processo de treinamento do mecanismo de identificação de *tweets* relacionados com alagamentos, combinação das informações contextuais ao nível de recurso e dos dados textuais ao nível de decisão, além do treinamento dos algoritmos de classificação da abordagem de Fusão Multimodal de modo híbrido, gasta-se cerca de **1 hora e 13 minutos** da máquina responsável pela realização dos testes.

Ademais, o modelo de **Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelos dados meteorológicos**, possui o intuito de combinar as informações textuais e geográficas da base de dados de treinamento com as informações resultantes do modelo de identificação de alagamentos, aliás instrui-se esse mecanismo de classificação de alagamentos a partir de informações climáticas na etapa de treinamento da abordagem de combinação de dados multimodais de modo tardio.

Dessa forma, os resultados do processo de ajuste de parâmetros dos algoritmos de classificação do modelo de combinação de dados heterogêneos de modo híbrido com foco na decisão conferida pelos dados climáticos são exibidos na Tabela 26. Sendo que, observa-se na Tabela 26 que o desempenho mais significativo resultante desse processo foi alcançado pelo DT, pois o impacto do aperfeiçoamento dos parâmetros é de 6,45, então conclui-se que os parâmetros

encontrados pelo módulo *GridSearchCV* são mais aderentes ao espaço amostral heterogêneo analisado do que os parâmetros definidos de forma padrão pela biblioteca *Scikit-Learn*. Por outro lado, também é possível notar na Tabela 26 que o resultado menos relevante obtido com o processo de ajuste de parâmetros foi alcançado pelo RF, pois o impacto da otimização foi de -5,58, logo infere-se que os parâmetros descobertos pelo *GridSearchCV* não são aderentes ao espaço amostral multimodal. Portanto, devido a este teste, emprega-se na etapa de treinamento do modelo de Fusão Multimodal de modo híbrido com foco na decisão promovida pelos dados climáticos os seguintes mecanismos de classificação com os parâmetros aperfeiçoados: SVM, DT e LR. Já quanto os algoritmos de classificação denominado RF e NB, empregam-se os parâmetros padrões obtidos com a biblioteca *Scikit-Learn*, porque o desempenho alcançado foi inferior após a otimização.

Tabela 26 – Resultados da otimização dos parâmetros utilizados pelos algoritmos de classificação empregados no modelo de Fusão Multimodal de modo híbrido com foco na decisão viabilizada pelos dados meteorológicos

Algoritmo	**Melhores parâmetros**	**Precisão com os parâmetros padrões**	**Precisão com os parâmetros otimizados**	**Impacto da otimização dos parâmetros**
SVM	‘C’: 1, ‘decision_function_shape’: ‘ovo’, ‘gamma’: 1, ‘kernel’: ‘rbf’, ‘shrinking’: True	0,7653	0,7653	0,00
RF	‘bootstrap’: True, ‘max_depth’: 80, ‘max_features’: 3, ‘min_samples_leaf’: 3, ‘min_samples_split’: 8, ‘n_estimators’: 100	0,8621	0,8063	-5,58
DT	‘criterion’: ‘gini’, ‘max_depth’: 2	0,7953	0,8599	6,45
LR	‘C’: 10, ‘multi_class’: ‘ovr’, ‘penalty’: ‘l2’, ‘solver’: ‘liblinear’	0,7332	0,7807	4,75
NB	‘var_smoothing’: 0.0187381	0,8000	0,7729	-2,71

Fonte: Elaborada pelo autor.

Assim, para a elaboração do mecanismo de **combinação de dados multimodais de modo híbrido com foco na decisão dos dados climáticos** realiza-se nesta etapa 50 testes de *ten-fold cross validation*, no qual são as diversas aplicações de algoritmos de ML com as várias categorias de conversão de informações simbólicas para numéricas (por exemplo, BOW, TF-IDF, *Word Embeddings* dos tipos *FastText* e *Word2Vec* das categorias *CBOW* e *Skip-gram* com 50 e 100 dimensões) no conjunto de dados denominado “Fusão Prévia” (Tabela 9). Desse modo, a seguir na Tabela 27 notam-se os melhores resultados do processo de treinamento desse modelo de Fusão Multimodal, inclusive, os resultados na íntegra dos testes deste experimento podem ser

observados no Apêndice G.

Tabela 27 – Melhores resultados do processo de treinamento do modelo de Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelos dados climáticos

Classificação	Algoritmo	Estratégia de transformação de dados textuais para numéricos	Precisão	Recall	F1-Score
1	DT	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,8721	0,8722	0,8722
2	DT	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,8721	0,8715	0,8718
3	RF	TF-IDF	0,8699	0,8652	0,8679
4	SVM	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,8684	0,8667	0,8675
5	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,8673	0,8611	0,8611

Fonte: Elaborada pelo autor.

Diante o exposto, é possível observar na Tabela 25 que o algoritmo de classificação responsável pelo melhor desempenho foi o **DT** com **0,8721** de precisão e utilizando a estratégia de transformação de dados baseada em ***Word Embeddings*** **do tipo** ***Word2Vec*** **da categoria** ***Skip-gram*** **com 100 dimensões** do NILC (HARTMANN *et al.*, 2017). Aliás, desde a execução do processo de treinamento do mecanismo de identificação de inundações a partir de dados climáticos, combinação das informações textuais e geográficas ao nível de recurso e dos dados meteorológicos ao nível de decisão, além do treinamento dos mecanismos de classificação da abordagem de Fusão Multimodal de modo híbrido, gasta-se cerca de **1 hora e 9 minutos** da máquina responsável pela realização dos testes.

Por último, nota-se que os processos de ajuste de parâmetros dos diferentes modelos de Fusão Multimodal em alguns casos particulares proporcionaram impactos extremamente relevantes na precisão dos mecanismos, por exemplo, a otimização dos parâmetros do SVM na abordagem de combinação de dados heterogêneos de modo prévio. Por outro lado, alguns ajustes de parâmetros não apresentaram um impacto no desempenho dos modelos de maneira significativa, por exemplo, o aperfeiçoamento dos parâmetros do RF no modelo de Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelos dados meteorológicos. Portanto, conclui-se que a realização da otimização dos parâmetros dos algoritmos de classificação trouxe diversos benefícios para as abordagens de combinação multimodal, pois em diversos casos o módulo *GridSearchCV* da biblioteca *Scikit-Learn* encontrou parâmetros mais aderentes ao espaço amostral multimodal analisado. Não menos importante, observa-se que as diversas abordagens de

combinação de dados multimodais treinados nesta etapa apresentaram resultados de treinamento significativos na tarefa de obtenção de SAW de alagamentos a partir de informações multimodais.

6.4 Comparação entre os modelos de Fusão Multimodal e averiguação do impacto da inclusão de dados contextuais na abordagem

O intuito deste experimento é proporcionar o treinamento de um modelo de identificação de possíveis vítimas de alagamentos a partir de dados unimodais, além de comparar o desempenho das abordagens de combinação de dados multimodais treinadas na Seção 6.3 e do modelo de classificação textual em um conjunto de dados inédito. Desse modo, avalia-se o impacto da inclusão de informações contextuais no desempenho dos modelos de combinação de informações heterogêneas e identifica-se abordagem mais precisa na tarefa de obtenção de SAW de inundações.

A princípio, para que seja possível o treinamento e teste da abordagem de identificação de possíveis vítimas de enchentes a partir de dados unimodais, logo é indispensável a execução da etapa de processamento textual, sendo que esta etapa seguiu a metodologia apresentada na Subsubseção 4.1.5.1 e usufruiu de estratégias de limpeza de dados e transformação de informações simbólicas para numéricas. Aliás, nesta etapa também avalia-se o impacto da utilização de diversas estratégias de transformação de dados (por exemplo, BOW, TF-IDF e *Word Embeddings*) quanto as métricas de avaliação definidas no Seção 6.1.

Logo após, existe a elaboração do modelo de classificação textual, no qual usufrui das informações textuais resultantes dos processos de limpeza e transformação de dados, além de aplicar a estratégia de classificação genérica exibida na Figura 26. Desse modo, o resultado do processo de ajuste de parâmetros dos algoritmos de ML da abordagem de classificação de *tweets* podem ser observados na Tabela 28.

Assim, na Tabela 28 observa-se que o impacto da execução do processo de otimização de parâmetros do NB é de 3,55, logo os parâmetros encontrados pelo *GridSearchCV* destacam-se pela aderência as informações textuais analisadas. No entanto, também é possível notar na Tabela 28 que o efeito do aperfeiçoamento dos parâmetros do RF é de -2,68, pois os parâmetros definidos como padrão pelo *Scikit-Learn* para esse algoritmo são mais favoráveis aos dados textuais analisados do que os encontrados pelo módulo *GridSearchCV* da biblioteca *Scikit-Learn*. Portanto, devido a este teste, utiliza-se na etapa de treinamento do modelo de classificação textual os seguintes algoritmos de classificação com os parâmetros otimizados: LR, NB. Já quanto aos demais mecanismos de classificação, emprega-se na etapa de treinamento os padrões obtidos com a biblioteca *Scikit-Learn*, porque os resultados obtidos foram menos precisos com os parâmetros aperfeiçoados.

Tabela 28 – Resultados da otimização dos parâmetros utilizados pelos algoritmos de classificação empregados no modelo de identificação de possíveis vítimas de alagamentos a partir de dados textuais

Algoritmo	**Melhores parâmetros**	**Precisão com os parâmetros padrões**	**Precisão com os parâmetros otimizados**	**Impacto da otimização dos parâmetros**
SVM	‘C’: 10, ‘decision_function_shape’: ‘ovo’, ‘gamma’: 0.001, ‘kernel’: ‘linear’, ‘shrinking’: True	0,6680	0,6576	-1,04
RF	‘bootstrap’: True, ‘max_depth’: 90, ‘max_features’: 3, ‘min_samples_leaf’: 3, ‘min_samples_split’: 8, ‘n_estimators’: 100	0,6666	0,6398	-2,68
DT	‘criterion’: ‘entropy’, ‘max_depth’: 44	0,6530	0,6469	-0,62
LR	‘C’: 25, ‘multi_class’: ‘ovr’, ‘penalty’: ‘l1’, ‘solver’: ‘liblinear’	0,6752	0,6789	0,38
NB	‘var_smoothing’: 1.0	0,6485	0,6840	3,55

Fonte: Elaborada pelo autor.

Ademais, os resultados mais expressivos do processo de treinamento dos algoritmos de ML do modelo de **classificação textual** podem ser observados na Tabela 29. Inclusive, nesta fase deste experimento foram executados 50 testes de *ten-fold cross validation*, onde são as aplicações dos diversos algoritmos de ML (por exemplo, RF, NB, SVM, DT, LR) para cada categoria de transformação de dados simbólicos para numéricos (por exemplo, BOW, TF-IDF, *Word Embeddings* do tipo *FastText* e *Word2Vec*) no espaço amostral textual de treinamento. Além disso, os resultados na íntegra dos testes desta etapa deste experimento podem ser observados no Apêndice H.

Dado o exposto, nota-se na Tabela 29 que o algoritmo de ML do modelo de classificação textual que obteve o melhor resultado na etapa de treinamento é o **NB** com **0,7007** de precisão e utilizando **BOW** como abordagem de transformação de dados simbólicos para numéricos. Aliás, desde a fase de limpeza de dados, transformação de informações textuais para numéricas e treinamento dos algoritmos de ML, o tempo gasto de processamento pela máquina utilizada para a execução dos testes desta etapa deste experimento foram de cerca de **1 hora e 17 minutos**.

Posteriormente a execução dos processos de treinamento dos modelos de combinação de dados multimodais e da abordagem de classificação unimodal, compara-se o desempenho das diversas abordagens de obtenção de SAW de inundações quanto as métricas de avaliação definidas no Seção 6.1 na base de dados inédita apresentada na Tabela 11. Desse modo, na Tabela 30 exibi-se os diversos resultados obtidos com a execução do processo de avaliação

Tabela 29 – Melhores resultados do processo de treinamento do modelo de identificação de possíveis vítimas de inundações a partir de dados textuais

Classificação	**Algoritmo**	**Estratégia de transformação de dados textuais para numéricos**	**Precisão**	**Recall**	**F1-Score**
1	NB	BOW	0,7007	0,6222	0,6535
2	RF	TF-IDF	0,6867	0,6833	0,6867
3	SVM	TF-IDF	0,6788	0,6760	0,6774
4	RF	BOW	0,6721	0,6667	0,6693
5	NB	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,6709	0,6167	0,6334

Fonte: Elaborada pelo autor.

dos modelos treinados nos experimentos da Seção 6.3 e Seção 6.4, inclusive as abordagens de captação de SAW de enchentes são apresentadas em ordem decrescente de precisão.

Tabela 30 – Comparação entre os modelos de combinação de dados multimodais e classificação textual

Classi-ficação	**Abordagem**	**Precisão**	**Recall**	**F1-Score**	**Tipos de dados**	**Tempo de exec (seg)**
1	**Fusão Multimodal de modo híbrido com foco na decisão viabilizada pelos dados climáticos**	**0,8470**	**0,8468**	**0,8468**	**Tweets e dados contextuais**	**0,234**
2	Fusão Multimodal de modo híbrido com foco na decisão viabilizada pelos dados textuais	0,8377	0,8288	0,8277	Tweets e dados contextuais	0,823
3	Fusão Multimodal de modo tardio	0,8377	0,8288	0,8277	Tweets e dados contextuais	0,945
4	Fusão Multimodal de modo prévio	0,8224	0,8198	0,8195	Tweets e dados contextuais	0,170
5	Modelo de classificação textual	0,6617	0,6351	0,6195	Tweets	0,196

Fonte: Elaborada pelo autor.

Dado o exposto, é possível observar na Tabela 30 que o "**Modelo de Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelas informações meteorológicas**" demonstrou maior precisão na tarefa de obtenção de SAW de enchentes que as demais abordagens, inclusive este modelo obteve **84,70%** de **precisão** ao ser avaliado na base de dados de validação (Tabela 11), além disso, esta abordagem apresenta um tempo de treinamento e de execução menor quando comparado com os demais modelos de combinação de dados heterogêneos. Aliás,

este modelo possui o intuito de explorar a correlação dos vetores de características dos *tweets* com os resultados da camada de decisão do mecanismo de identificação de inundações.

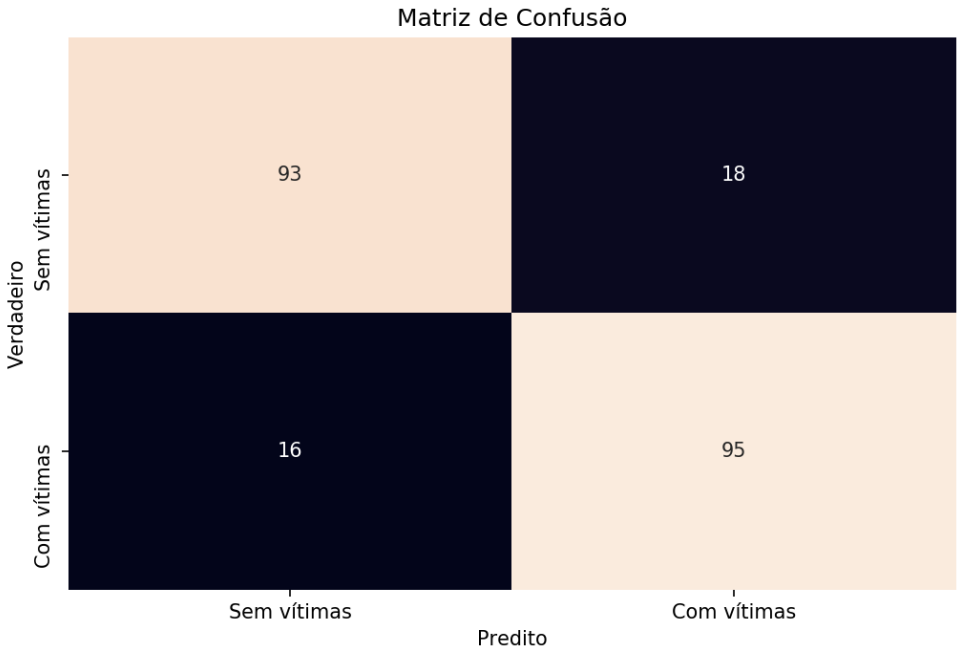
Já o "Modelo de Fusão Multimodal de modo tardio" possui o desempenho inferior quando comparado com o primeiro colocado, pois ao combinar as informações textuais com os dados contextuais há a simplificação da dinâmica intermodal (BRUIJN *et al.*, 2020). Além disso, observa-se que a abordagem de combinação de dados heterogêneos de modo tardio possui o tempo de treinamento e de execução demasiado quando comparado com as demais abordagens de combinação de informações multimodais. Ademais, de acordo com Lopes (2015) o "Modelo de Fusão Multimodal de modo prévio" possui uma grande probabilidade de gerar *overfitting* e transtornos relacionados com a sincronização das informações heterogêneas, por isso esse modelo é menos preciso que o primeiro colocado. Por último, o "Modelo de Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelos dados textuais" transporta para a camada de combinação de dados multimodais os vieses encontrados na camada de classificação textual, por esse motivo que este modelo tende a ser menos preciso do que a abordagem de combinação de dados multimodais de modo híbrido com foco na decisão viabilizada pelas informações climáticas.

Além disso, nota-se na Figura 39 a matriz de confusão do "**Modelo de Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelas informações meteorológicas**", no qual foi elaborada a partir dos rótulos do conjunto de dados de validação (Tabela 11) e dos valores preditos pela abordagem de combinação de dados multimodais ao ser executada nessa base de dados inédita. Inclusive, também é possível observar na Figura 39 que a abordagem de combinação de dados midiáticos de modo híbrido não gerou *overfitting*, visto que se observa que algumas vezes o modelo erra a classificação de alguns exemplos do espaço amostral (por exemplo, a classificação de possíveis vítimas de alagamentos como não possíveis vítimas de alagamentos).

Conclui-se que, o modelo de Fusão Multimodal de modo híbrido com foco na decisão viabilizada pelos dados climáticos é o mais preciso dentre os modelos de combinação de dados multimodais testados, além de que esta abordagem não gerou *overfitting* e o tempo de treinamento e execução é menor que uma parcela dos modelos de combinação de dados multimodais analisados. Aliás, o tempo de execução do modelo de combinação de dados heterogêneos de modo tardio é de cerca de quatro vezes superior ao do modelo que é o primeiro colocado da Tabela 30, pois no modelo que se destaca dentre os demais devido à precisão há somente duas etapas, ou seja, a identificação de alagamentos e a combinação de dados multimodais, já na abordagem que é a segunda da classificação existem três etapas, ou seja, a identificação de *tweets* relacionados com inundações, a classificação de dados climáticos quanto ao relacionamento com alagamentos e a Fusão Multimodal.

Por último, a **inclusão de dados contextuais** nas abordagens de Fusão Multimodal indica a **melhora de 18,53%** na tarefa de obtenção de SAW de alagamentos da cidade de São

Figura 39 – Matriz de confusão do Modelo de Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelas informações meteorológicas



Fonte: Elaborada pelo autor.

Paulo, ou seja, a diferença entre o desempenho do modelo de Fusão Multimodal de modo híbrido com foco na decisão viabilizada pelos dados meteorológicos e o modelo de classificação textual. Desse modo, abordagens que utilizam dados multimodais (por exemplo, dados textuais, geográficos e climáticos) tendem a ter resultados mais promissores do que modelos que utilizam informações unimodais (por exemplo, informações textuais).

6.5 Considerações Finais

Este capítulo apresentou as métricas de avaliação e os resultados dos seguintes experimentos realizados nesta pesquisa: Descoberta das regiões propícias ao acontecimento de alagamentos; Treinamento e avaliação dos modelos de Fusão Multimodal; Comparação entre os modelos de Fusão Multimodal e averiguação do impacto da inclusão de dados contextuais na abordagem. Desse modo, no primeiro experimento há a identificação das áreas de propícias ao acontecimento de inundações na cidade de São Paulo, além da comprovação de que abordagens empíricas para a definição da distância máxima de criação de grupos produzem regiões de enchentes mais bem definidas do que geo estatísticas. Já no segundo o experimento, observou-se que o processo de ajuste de parâmetros dos algoritmos de ML proporcionou impactos extremamente relevantes para esta pesquisa, além de que as diversas abordagens de combinação de dados multimodais atingiram resultados significativos na etapa de treinamento. Por último, no terceiro experimento constatou-se que a abordagem de combinação de dados heterogêneos mais precisa na tarefa de

captar SAW de alagamentos foi o modelo de Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelos dados climáticos (0,8740 de precisão), aliás houve a demonstração de que combinar informações contextuais com textuais aumenta a precisão na tarefa de detecção de possíveis vítimas de enchentes (18,53%).

CAPÍTULO

7

CONCLUSÃO

Neste trabalho foi apresentado uma abordagem de Fusão Multimodal, no qual pode detectar possíveis vítimas de alagamentos da cidade de São Paulo e auxiliar a etapa de resposta da GD. Sendo que, as informações contextuais utilizadas pelos modelos de Fusão Multimodal explorados nesta dissertação são: dados textuais, informações meteorológicas e ocorrências históricas de enchentes.

Nesta pesquisa é demonstrado que fundir informações textuais e dados climáticos com informações geográficas advindas de regiões propícias ao acontecimento de inundações, amplia-se consideravelmente a precisão da detecção de possíveis vítimas de inundações (18,53%), pois os modelos de identificação que utilizam somente características textuais possuem a tendência de terem um desempenho inferior se comparados com os modelos multimodais. Além disso, dentre os modelos de combinação de informações heterogêneas explorados nesta pesquisa, o mecanismo computacional com maior precisão na execução da tarefa de obtenção de SAW de alagamentos foi o modelo de Fusão Multimodal do tipo híbrido com foco na decisão proporcionada pelos dados climáticos (84,70%), visto que se utiliza nesta estratégia de Fusão Multimodal a combinação dos vetores de características dos *tweets* e das informações geográficas com as predições do mecanismo de classificação de alagamentos, assim evita-se *overfitting* e uma simplificação demasiada da dinâmica intermodal.

Ademais, nesta dissertação exploram-se diversas estratégias de agrupamento de ocorrências históricas de inundações e de definição de distância máxima de criação de *clusters*, dessa forma nota-se que o algoritmo hierárquico utilizado para descobrir as regiões propícias ao acontecimento de alagamentos da cidade de São Paulo teve um desempenho superior que as demais estratégias de *clustering* baseadas em densidade, ou seja, o melhor resultado encontrado de *Silhouette* no experimento de descoberta das áreas de alagamentos foi pelo algoritmo *Agglomerative Clustering* com o critério de ligação de grupos do tipo *Average* (0,4788). Inclusive, as distâncias máximas de criação de grupos utilizadas pelos mecanismos de *clustering* que

foram definidas de maneira empírica obtiveram desempenhos superiores do que a estatística (Semivariograma), pois as enchentes são desastres naturais que ocorrem devido à intervenção inadequada dos seres humanos no meio ambiente, já o Semivariograma é uma técnica Geoestatística empregada em cenários naturais que não sofreram as intervenções dos seres humanos (por exemplo, jazidas de ouro) (YIN; LI, 2001).

Além disso, o *software* desenvolvido neste trabalho é capaz de detectar SAW de enchentes e auxiliar a etapa de resposta da GD em tempo real, assim a DC, os bombeiros e as ONGs podem utilizar as informações processadas por esta plataforma para socorrer as vítimas dos desastres naturais da cidade de São Paulo. Inclusive, esta plataforma possui diversas funcionalidades, como: detecção de possíveis vítimas de enchentes, identificação de áreas de inundações ativas e análise de dados históricos e em tempo real. Não menos importante, o *software* foi construído baseado em uma arquitetura de multicamadas (ou seja, *Urls*, *Models*, *Views*, *Templates*, *Persistência* e Detecção de vítimas de alagamentos), onde cada camada possui funcionalidades específicas, por exemplo, a camada de *Models* contém os campos e as maneiras de interação com as informações persistidas.

Dado o exposto, esta dissertação elabora abordagens de combinação de dados heterogêneos de modo híbrido e explora estratégias de Fusão Multimodal clássicas (por exemplo, abordagem prévia e tardia), além disso, efetua a comparação entre os modelos de combinação de dados multimodais desenvolvidos e realiza a investigação do impacto da inclusão de informações contextuais nas abordagens de Fusão Multimodal ao comparar os seguintes modelos na tarefa de identificar possíveis vítimas de inundações da cidade de São Paulo: modelo de Fusão Multimodal de modo híbrido com foco na decisão proporcionada pelas informações meteorológicas e modelo de classificação unimodal. Assim, conclui-se que estratégias multimodais para auxílio na fase de resposta da GD são mais eficazes que as unimodais. Aliás, os modelos de combinação de dados multimodais desenvolvidos nesta dissertação são facilmente estendidos para a aquisição de SAW de diferentes desastres naturais, em diversas regiões do globo terrestre e a partir de mensagens de redes sociais escritas em outros idiomas.

7.1 Principais Contribuições

A seguir apresentam-se as principais vantagens que este trabalho acrescenta para a área de Fusão Multimodal como instrumento de apoio a GD:

- Análise e síntese dos recentes trabalhos correlatos encontrados na literatura, sendo que as áreas selecionadas são: Gestão de Desastres, Fusão de Dados, Mineração de Texto, Análise de Redes Sociais e Aprendizado de Máquina. Aliás, esta contribuição pode ser observada no Capítulo 3.

- Elaboração de uma abordagem para a descoberta das possíveis áreas de alagamentos da cidade de São Paulo. Ademais, esta contribuição pode ser notada na Seção 6.2.

- Constata-se que as abordagens empíricas de definição da distância máxima de criação de agrupamentos produzem áreas de alagamentos mais bem formadas em comparação com a geo estatística. Inclusive, observa-se esta contribuição na Seção 6.2.

- Criação de uma base de dados inédita para o treinamento e teste dos modelos de detecção de possíveis vítimas de enchentes, no qual é constituída de dados textuais, climáticos e geográficos. Além disso, nota-se a construção desta base de dados na Subseção 4.1.6, além de ser disponibilizada via *Github*[1].

- Elaboração de um modelo de Fusão Multimodal do tipo híbrido capaz de obter SAW de alagamentos e auxiliar a etapa de resposta da GD. Aliás, esta contribuição é observada na Subseção 4.1.7 e Seção 6.3.

- Constata-se que o modelo de Fusão Multimodal do tipo híbrido com foco na decisão viabilizada pelos dados climáticos é mais preciso na tarefa de detectar possíveis vítimas de inundações do que as estratégias de combinação de dados heterogêneos dos seguintes tipos: prévio, tardio e híbrido com foco na decisão proporcionada pelas informações textuais. Inclusive, esta contribuição é notada na Seção 6.4.

- Constata-se que combinar dados multimodais aumenta a precisão na tarefa de detecção de vítimas de enchentes. Ademais, observa-se esta contribuição na Seção 6.4.

- Desenvolvimento de uma plataforma de reconhecimento de possíveis vítimas de alagamentos em tempo real. Além disso, é possível notar esta contribuição no Capítulo 5.

7.2 Trabalhos futuros

Nesta dissertação foi desenvolvida uma abordagem de combinação de informações multimodais com o intuito de auxiliar a etapa de resposta da GD, além da criação de um *software* capaz de detectar possíveis vítimas de alagamentos em tempo real, assim surgiram diversas ideias que possibilitam o aperfeiçoamento desses mecanismos computacionais, desse modo serão listadas a seguir:

- Utilizar outras abordagens para a definição da distância máxima de criação de *clusters*, por exemplo, a estatística clássica;

[1] url: <https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/train_test/data>

- Aplicar outros algoritmos de agrupamento nas ocorrências históricas de alagamentos para descobrir as regiões propícias ao acontecimento de enchentes da cidade São Paulo, por exemplo, redes neurais;
- Utilizar abordagens de NLP mais recentes para a execução da conversão de informações simbólicas para numéricas, por exemplo, DistilBERT (SANH *et al.*, 2019);
- Identificar *tweets* relacionados com outras fases da GD, por exemplo, mitigação, preparação e recuperação. Desse modo, é possível expandir a atuação do modelo de Fusão Multimodal apresentado nesta dissertação, além de auxiliar outras fases da GD;
- Propor uma abordagem de identificação de possíveis vítimas de alagamentos baseada em CNN, além de comparar os resultados produzidos por essa abordagem com as estratégias de combinação de dados multimodais híbridas apresentadas nesta dissertação;
- Realizar a validação estatística dos modelos de combinação de informações multimodais, assim é possível identificar se existe diferença estatística entre os algoritmos de ML empregados;
- Utilizar o mecanismo computacional de armazenamento de dados em memória chamado Redis[2] para guardar o *cache* do serviço de análise de dados históricos da plataforma de detecção de possíveis vítimas de inundações, assim evita-se o reprocessamento das informações e possibilita o armazenamento dos dados por um período limitado aos usuários ativos do *software*;
- Realizar experimentos de usabilidade e de carga no *software* de detecção de vítimas de enchentes. Dessa maneira, possibilita-se a confecção de interfaces mais agradáveis e úteis aos usuários, além da utilização de serviços de nuvem mais adequados.

7.3 Atividades Acadêmicas e Complementares

7.3.1 Disciplinas Concluídas e Atividades Realizadas

Este discente, durante o ano de 2018, concluiu as seguintes disciplinas do programa de pós-graduação em Ciências da Computação e Matemática Computacional do ICMC-USP: Ciência de Dados; Redes Neurais; Mineração de Dados Não Estruturados; Metodologia de Pesquisa Científica em Computação; Metodologia de Pesquisa Científica em Bancos de Dados e Imagens; Tópicos em Inteligência Artificial; Preparação Pedagógica. Ademais, durante o ano de 2019, o aluno foi o Vice-Representante Discente da Comissão de Pós-Graduação do ICMC-USP e Vice-Representante Discente da Comissão de Relações Internacionais do ICMC-USP, além disso, o aluno realizou os seguintes estágios do Programa de Aperfeiçoamento de Ensino (PAE)

[2] url: <https://redis.io/>

do ICMC-USP sob supervisão do professor Dr. Jó Ueyama: Redes de Computadores e Redes Móveis, sendo que o primeiro estágio o aluno realizou como voluntário, já o segundo como bolsista.

7.3.2 Produção Científica

7.3.2.1 Artigos Publicados

International Journal **(Qualis A1)**: **Costa, T. A. G.**, Meneguette, R. I., Ueyama, J. "*Providing a greater precision of Situational Awareness of urban floods through Multimodal Fusion*". In: *Expert Systems with Applications*, 2021.

Workshop: Andrade, S. C.; Estrada, C. R.; **Costa, T. A. G.**; Ueyama, J.; Delbem, A. C. B.; Albuquerque, J. P. "*Situational awareness in social media: lessons learned using information entropy in flood risk management*". In: Segundo Workshop NUVEM, 2018, Santo André. Anais do segundo Workshop NUVEM, 2018.

7.3.3 Produção Técnica

Registro de software: Ueyama, J.; **Costa, T. A. G.** "SOFTWARE PARA A IDENTIFICAÇÃO DE VÍTIMAS DE ALAGAMENTOS A PARTIR DE TWEETS PÚBLICOS E DADOS METEOROLÓGICOS". Número do registro: BR512020001640-5, data de registro: 24/04/2020, Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial, 2020.

7.3.4 Revisão de Artigos Científicos

- Seminário Integrado de *Software* e *Hardware* - SEMISH 2021;
- 38º Simpósio Brasileiro de Redes de Computadores - 38º SBRC 2020;
- IV Workshop de Computação Urbana - IV COURB 2020;
- III Workshop de Computação Urbana - III COURB 2019;
- 46º Seminário Integrado de *Software* e *Hardware* - SEMISH 2019;
- 11ª Conferência Latino-Americana de Comunicações - LATINCOM 2019.

REFERÊNCIAS

AGGARWAL, C. C.; ZHAI, C. **Mining text data**. [S.l.]: Springer Science & Business Media, 2012. Citado na página 44.

AGHI, R.; MEHTA, S.; CHAUHAN, R.; CHAUDHARY, S.; BOHRA, N. A comprehensive comparison of sql and mongodb databases. **International Journal of Scientific and Research Publications**, Citeseer, v. 5, n. 2, p. 1–3, 2015. Citado na página 83.

AGUIAR, E. J. de; FAIÇAL, B. S.; UEYAMA, J.; SILVA, G. C.; MENOLLI, A. Análise de sentimento em redes sociais para a língua portuguesa utilizando algoritmos de classificação. In: **Anais do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**. Porto Alegre, RS, Brasil: SBC, 2018. ISSN 2177-9384. Disponível em: <https://sol.sbc.org.br/index.php/sbrc/article/view/2430>. Citado nas páginas 45 e 90.

ALBUQUERQUE, J. P. de; HERFORT, B.; BRENNING, A.; ZIPF, A. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. **International Journal of Geographical Information Science**, v. 29, n. 4, p. 667–689, 2015. Disponível em: <https://doi.org/10.1080/13658816.2014.996567>. Citado nas páginas 31, 32 e 85.

ALMEIDA, F.; XEXÉO, G. Word embeddings: A survey. **arXiv preprint arXiv:1901.09069**, 2019. Citado na página 47.

ALQHTANI, S. M.; LUO, S.; REGAN, B. Multimedia data fusion for event detection in twitter by using dempster-shafer evidence theory. **World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering**, v. 9, n. 12, p. 2234–2238, 2015. Citado nas páginas 42, 65, 66, 73 e 74.

AMARAL, R. d.; GUTJAHR, M. R. Desastres naturais (série cadernos de educação ambiental, 8). **São Paulo: IG/SMA**, 2011. Citado na página 29.

ANDRADE, S. C.; RESTREPO-ESTRADA, C.; COSTA, T. A. G.; UEYAMA, J.; DELBEM, A. C. B.; ALBUQUERQUE, J. P. Situational awareness in social media: lessons learnedusing information entropy in flood risk management. **2º Workshop NUVEM**, Universidade Federal do ABC, p. 1–4, 2018. Citado na página 50.

ANDRADE, S. C. d. **Mining of rainfall patterns from social media for supporting flood risk management**. Tese (Doutorado) — Universidade de São Paulo, 2020. Citado na página 85.

ANDRADE, S. C. de; DEGROSSI, L. C.; ESTRADA, C. R.; DELBEM, A. C.; ALBUQUERQUE, J. P. de. Does keyword noise change over space and time? a case study of social media messages. In: **GEOINFO**. [S.l.: s.n.], 2018. p. 116–121. Citado nas páginas 82 e 85.

ANDRADE, S. C. de; RESTREPO-ESTRADA, C.; DELBEM, A. C.; MENDIONDO, E. M.; ALBUQUERQUE, J. P. de. Mining rainfall spatio-temporal patterns in twitter: a temporal approach. In: SPRINGER. **The Annual International Conference on Geographic Information Science**. [S.l.], 2017. p. 19–37. Citado nas páginas 41, 79, 80 e 85.

ANDRADE, S. C. de; RESTREPO-ESTRADA, C.; NUNES, L. H.; RODRIGUEZ, C. A. M.; ESTRELLA, J. C.; DELBEM, A. C. B.; ALBUQUERQUE, J. Porto de. A multicriteria optimization framework for the definition of the spatial granularity of urban social media analytics. **International Journal of Geographical Information Science**, Taylor & Francis, p. 1–20, 2020. Citado na página 82.

ANKERST, M.; BREUNIG, M. M.; KRIEGEL, H.-P.; SANDER, J. Optics: ordering points to identify the clustering structure. **ACM Sigmod record**, ACM New York, NY, USA, v. 28, n. 2, p. 49–60, 1999. Citado nas páginas 55, 56 e 74.

ASHKTORAB, Z.; BROWN, C.; NANDI, M.; CULOTTA, A. Tweedr: Mining twitter to inform disaster response. In: **ISCRAM**. [S.l.: s.n.], 2014. p. 269–272. Citado na página 119.

ASSIS, L. F. F. G. de; ALBUQUERQUE, J. P. de; HERFORT, B.; STEIGER, E.; HORITA, F. E. A. Geographical prioritization of social network messages in near real-time using sensor data streams: an application to floods. **Revista Brasileira de Cartografia**, v. 68, n. 6, 2016. Citado na página 85.

ATREY, P. K.; HOSSAIN, M. A.; SADDIK, A. E.; KANKANHALLI, M. S. Multimodal fusion for multimedia analysis: a survey. **Multimedia systems**, Springer, v. 16, n. 6, p. 345–379, 2010. Citado nas páginas 40, 42, 43, 95 e 96.

AVVENUTI, M.; CRESCI, S.; MARCHETTI, A.; MELETTI, C.; TESCONI, M. Ears (earthquake alert and report system) a real time decision support system for earthquake crisis management. In: **Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2014. p. 1749–1758. Citado na página 31.

BAHARIN, S. S. K.; SHIBGHATULLAH, A. S.; OTHMAN, Z. Disaster management in malaysia: An application framework of integrated routing application for emergency response management system. In: IEEE. **2009 International Conference of Soft Computing and Pattern Recognition**. [S.l.], 2009. p. 716–719. Citado nas páginas 30 e 36.

BARONI, M.; DINU, G.; KRUSZEWSKI, G. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. [S.l.: s.n.], 2014. v. 1, p. 238–247. Citado na página 47.

BATISTA, G. E. d. A. P. *et al.* **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese (Doutorado) — Universidade de São Paulo, 2003. Citado na página 45.

BIRANT, D.; KUT, A. St-dbscan: An algorithm for clustering spatial–temporal data. **Data & Knowledge Engineering**, Elsevier, v. 60, n. 1, p. 208–221, 2007. Citado na página 55.

BOETTCHER, A.; LEE, D. Eventradar: A real-time local event detection scheme using twitter stream. In: IEEE. **2012 IEEE International Conference on Green Computing and Communications**. [S.l.], 2012. p. 358–367. Citado na página 50.

BORAH, B.; BHATTACHARYYA, D. An improved sampling-based dbscan for large spatial databases. In: IEEE. **International Conference on Intelligent Sensing and Information Processing, 2004. Proceedings of**. [S.l.], 2004. p. 92–96. Citado nas páginas 55 e 74.

BOTEGA, L. C. **Modelo de Fusão Dirigido por Humanos e Ciente de Qualidade de Informação**. Tese (Doutorado) — Universidade Federal de São Carlos, 2016. Citado nas páginas 37, 38, 39 e 40.

BRAGA, A. de P. **Redes neurais artificiais: teoria e aplicações**. LTC Editora, 2007. ISBN 9788521615644. Disponível em: <https://books.google.com.br/books?id=R-p1GwAACAAJ>. Citado na página 52.

BROUWER, T.; EILANDER, D.; LOENEN, A. V.; BOOIJ, M. J.; WIJNBERG, K. M.; VERKADE, J. S.; WAGEMAKER, J. Probabilistic flood extent estimates from social media flood observations. **Natural Hazards & Earth System Sciences**, v. 17, n. 5, 2017. Citado na página 31.

BRUIJN, J. A. de; MOEL, H. de; JONGMAN, B.; WAGEMAKER, J.; AERTS, J. C. Taggs: grouping tweets to improve global geoparsing for disaster response. **Journal of Geovisualization and Spatial Analysis**, Springer, v. 2, n. 1, p. 2, 2018. Citado na página 31.

BRUIJN, J. A. de; MOEL, H. de; WEERTS, A. H.; RUITER, M. C. de; BASAR, E.; EILANDER, D.; AERTS, J. C. Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network. **Computers & Geosciences**, Elsevier, p. 104485, 2020. Citado nas páginas 31, 32, 41, 42, 43, 45, 50, 71, 72, 73, 74, 75, 76, 77, 87, 90, 95, 99, 119 e 140.

CAMARGO, E. C. G. **Geoestatística: fundamentos e aplicações**. [S.l.: s.n.], 1998. Citado nas páginas 61 e 62.

CASTRO, A. L. C. d.; CALHEIROS, L. B.; CUNHA, M. I. R.; BRINGEL, M. L. N. d. C. **Manual de desastres: desastres naturais**. [S.l.]: Ministério da Integração Nacional, 1996. Citado na página 29.

CHEN, M.; CHEN, S.-C.; SHYU, M.-L.; WICKRAMARATNA, K. Semantic event detection via multimodal data mining. **IEEE Signal Processing Magazine**, IEEE, v. 23, n. 2, p. 38–46, 2006. Citado na página 93.

CIGAGNA, C.; BONOTTO, D. M.; STURARO, J. R.; CAMARGO, A. F. M. Geostatistical techniques applied to mapping limnological variables and quantify the uncertainty associated with estimates. **Acta Limnologica Brasiliensia**, SciELO Brasil, v. 27, n. 4, p. 421–430, 2015. Citado na página 124.

CLARK, I. **Practical geostatistics**. [S.l.]: Applied Science Publishers London, 1979. v. 3. Citado nas páginas 59 e 60.

COLLOBERT, R.; WESTON, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: ACM. **Proceedings of the 25th international conference on Machine learning**. [S.l.], 2008. p. 160–167. Citado na página 47.

CREPALDI, P. G.; AVILA, R. N. P.; OLIVEIRA, J. P. M. de; RODRIGUES, P. R.; MARTINS, R. L. Um estudo sobre a árvore de decisão e sua importância na habilidade de aprendizado. **Revista Eletrônica do Instituto de Ensino Superior de Londrina**, 2011. Citado na página 50.

CROOKS, A.; CROITORU, A.; STEFANIDIS, A.; RADZIKOWSKI, J. # earthquake: Twitter as a distributed sensor system. **Transactions in GIS**, Wiley Online Library, v. 17, n. 1, p. 124–147, 2013. Citado na página 31.

CUTTER, S. L.; EMRICH, C. Are natural hazards and disaster losses in the us increasing? **EOS, Transactions American Geophysical Union**, Wiley Online Library, v. 86, n. 41, p. 381–389, 2005. Citado na página 35.

DEGROSSI, L. C. **Uma abordagem para obtenção e disponibilização em tempo real de informações geográficas voluntárias no contexto de gestão de risco de inundação**. Dissertação (Mestrado) — Universidade de São Paulo, 2015. Citado na página 36.

DOBSON, A. J.; BARNETT, A. G. **An introduction to generalized linear models**. [S.l.]: CRC press, 2018. Citado na página 54.

DOMINGOS, P.; PAZZANI, M. On the optimality of the simple bayesian classifier under zero-one loss. **Machine learning**, Springer, v. 29, n. 2-3, p. 103–130, 1997. Citado na página 51.

EARLE, P. S.; BOWDEN, D. C.; GUY, M. Twitter earthquake detection: earthquake monitoring in a social world. **Annals of Geophysics**, v. 54, n. 6, 2012. Citado na página 31.

ENDSLEY, M. R. Design and evaluation for situation awareness enhancement. In: SAGE PUBLICATIONS SAGE CA: LOS ANGELES, CA. **Proceedings of the Human Factors Society annual meeting**. [S.l.], 1988. v. 32, n. 2, p. 97–101. Citado nas páginas 31, 37, 38 e 39.

____. Designing for situation awareness in complex systems. In: **Proceedings of the Second International Workshop on symbiosis of humans, artifacts and environment**. [S.l.: s.n.], 2001. p. 1–14. Citado na página 38.

____. **Designing for situation awareness: An approach to user-centered design**. [S.l.]: CRC press, 2016. Citado na página 38.

ESPEJO, T. M. S. **Interference Due to Rain in Urban Environments for Millimeters Waves**. Tese (Doutorado) — Pontifícia Universidade Católica do Rio de Janeiro, 2016. Citado na página 100.

FACEBOOK. **FastText: Library for efficient text classification and representation learning.** 2020. Acessado 09 nov. 2020. Disponível em: <https://fasttext.cc/docs/en/support.html>. Citado na página 47.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. *et al.* **Inteligência Artificial: Uma abordagem de aprendizado de máquina**. [S.l.: s.n.], 2011. v. 2. 192 p. Citado nas páginas 41, 45, 46, 48, 49, 50, 51, 52, 53, 56, 57, 74, 85, 86, 95, 98, 102, 120 e 121.

FÉLIX, V. B.; JÚNIOR, O. A. G.; ROSSONI, D. F.; HENRIQUES, M. J. Estimadores de semivariância: Uma revisão. **Ciência e Natura**, Universidade Federal de Santa Maria, v. 38, n. 3, p. 1157–1167, 2016. Citado na página 59.

FENG, Y.; SESTER, M. Extraction of pluvial flood relevant volunteered geographic information (VGI) by deep learning from user generated texts and photos. **ISPRS Int. J. Geo-Information**, v. 7, n. 2, p. 39, 2018. Disponível em: <https://doi.org/10.3390/ijgi7020039>. Citado nas páginas 31, 32, 50, 68, 69, 70, 72, 73, 74, 75, 77, 100 e 121.

FIGUEIRA, C. V. **Modelos de regressão logística**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, 2006. Citado na página 53.

FOUNDATION, D. S. **Despachante de URL**. Foundation, Django Software, 2021. Acessado em: 18/02/2021. Disponível em: <https://docs.djangoproject.com/pt-br/3.0/topics/http/urls/>. Citado na página 109.

____. **Escrevendo "views"**. Foundation, Django Software, 2021. Acessado em: 18/02/2021. Disponível em: <https://docs.djangoproject.com/pt-br/3.0/topics/http/views/>. Citado na página 109.

____. **Models**. Foundation, Django Software, 2021. Acessado em: 18/02/2021. Disponível em: <https://docs.djangoproject.com/pt-br/3.0/topics/db/models/>. Citado na página 109.

____. **Templates**. Foundation, Django Software, 2021. Acessado em: 24/02/2021. Disponível em: <https://docs.djangoproject.com/pt-br/3.0/topics/templates/>. Citado na página 110.

FRIEDL, M. A.; BRODLEY, C. E. Decision tree classification of land cover from remotely sensed data. **Remote sensing of environment**, Elsevier, v. 61, n. 3, p. 399–409, 1997. Citado nas páginas 50 e 51.

GONÇALVES, E. B.; GOUVÊA, M. A.; MANTOVANI, D. M. N. Análise de risco de crédito com o uso de regressão logística. **Revista Contemporânea de Contabilidade**, v. 10, n. 20, p. 139–160, 2013. Citado nas páginas 54 e 55.

GOOGLE. **word2vec: Tool for computing continuous distributed representations of words.** 2013. Acessado 09 nov. 2020. Disponível em: <https://code.google.com/archive/p/word2vec/>. Citado na página 47.

HADDAD, E. A.; TEIXEIRA, E. Economic impacts of natural disasters in megacities: The case of floods in são paulo, brazil. **Habitat International**, Elsevier, v. 45, p. 106–113, 2015. Citado nas páginas 29 e 30.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; RODRIGUES, J.; ALUISIO, S. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. **arXiv preprint arXiv:1708.06025**, 2017. Citado nas páginas 89, 90, 131, 134 e 136.

HEYDON, A.; NAJORK, M. Mercator: A scalable, extensible web crawler. **World Wide Web**, Springer, v. 2, n. 4, p. 219–229, 1999. Citado na página 82.

HO, T. K. Random decision forests. In: IEEE. **Proceedings of 3rd international conference on document analysis and recognition**. [S.l.], 1995. v. 1, p. 278–282. Citado na página 50.

HORITA, F. E.; ALBUQUERQUE, J. P. de; DEGROSSI, L. C.; MENDIONDO, E. M.; UEYAMA, J. Development of a spatial decision support system for flood risk management in brazil that combines volunteered geographic information with wireless sensor networks. **Computers & Geosciences**, Elsevier, v. 80, p. 84–94, 2015. Citado nas páginas 30 e 37.

HORITA, F. E. A. **An approach for improving decision-making with heterogeneous geospatial big data: an application using spatial decision support systems and volunteered geographic information to disaster management**. Tese (Doutorado) — Universidade de São Paulo, 2017. Citado na página 36.

HUANG, C.; DAVIS, L.; TOWNSHEND, J. An assessment of support vector machines for land cover classification. **International Journal of remote sensing**, Taylor & Francis, v. 23, n. 4, p. 725–749, 2002. Citado na página 53.

HUANG, Q.; XIAO, Y. Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. **ISPRS International Journal of Geo-Information**, Multidisciplinary Digital Publishing Institute, v. 4, n. 3, p. 1549–1568, 2015. Citado nas páginas 50, 66, 67, 72, 73, 74, 75, 85 e 99.

HUGHES, D.; UEYAMA, J.; MENDIONDO, E.; MATTHYS, N.; HORRÉ, W.; MICHIELS, S.; HUYGENS, C.; JOOSEN, W.; MAN, K. L.; GUAN, S.-U. A middleware platform to support river monitoring using wireless sensor networks. **Journal of the Brazilian Computer Society**, Springer, v. 17, n. 2, p. 85–102, 2011. Citado na página 30.

IBGE. **Censo Demográfico**. [S.l.]: Instituto Brasileiro de Geografia e Estatística, 2010. Citado nas páginas 36 e 79.

ISAAKS, E. H.; SRIVASTAVA, R. M. **An introduction to applied geostatistics**. [S.l.], 1989. Citado nas páginas 58, 59, 61, 62 e 122.

JOULIN, A.; GRAVE, E.; BOJANOWSKI, P.; DOUZE, M.; JÉGOU, H.; MIKOLOV, T. Fasttext.zip: Compressing text classification models. **arXiv preprint arXiv:1612.03651**, 2016. Citado na página 47.

KELLY, S.; ZHANG, X.; AHMAD, K. Mining multimodal information on social media for increased situational awareness. **Proceedings of the 14th ISCRAM Conference**, p. 613–622, 2017. Citado nas páginas 67, 68 e 75.

KISHI, R. M. **Fusão de informação multimodal por detecção de correlação para tarefas de análise de vídeo**. Tese (Doutorado) — Universidade de São Paulo, 2020. Citado na página 42.

KLOMP, J. Economic development and natural disasters: A satellite data analysis. **Global Environmental Change**, Elsevier, v. 36, p. 67–88, 2016. Citado na página 35.

KOBIYAMA, M.; MENDONÇA, M.; MORENO, D. A.; MARCELINO, I.; MARCELINO, E. V.; GONÇALVES, E. F.; BRAZETTI, L. L.; GOERL, R. F.; MOLLERI, G. S.; RUDORFF, F. d. M. **Prevenção de desastres naturais: conceitos básicos**. [S.l.]: Organic Trading Curitiba, 2006. Citado nas páginas 30 e 35.

KOKAR, M. M.; ENDSLEY, M. R. Situation awareness and cognitive modeling. **IEEE Intelligent Systems**, IEEE, v. 27, n. 3, p. 91–96, 2012. Citado na página 38.

KONONENKO, I. Semi-naive bayesian classifier. In: SPRINGER. **European Working Session on Learning**. [S.l.], 1991. p. 206–219. Citado na página 51.

KRIGE, D. G. A statistical approach to some basic mine valuation problems on the witwatersrand. **Journal of the Southern African Institute of Mining and Metallurgy**, Southern African Institute of Mining and Metallurgy, v. 52, n. 6, p. 119–139, 1951. Citado na página 58.

KRIPPENDORFF, K. Reliability in content analysis: Some common misconceptions and recommendations. **Human communication research**, Wiley Online Library, v. 30, n. 3, p. 411–433, 2004. Citado na página 87.

KRYVASHEYEU, Y.; CHEN, H.; OBRADOVICH, N.; MORO, E.; HENTENRYCK, P. V.; FOWLER, J.; CEBRIAN, M. Rapid assessment of disaster damage using social media activity. **Science advances**, American Association for the Advancement of Science, v. 2, n. 3, p. e1500779, 2016. Citado na página 31.

KWAK, H.; LEE, C.; PARK, H.; MOON, S. What is twitter, a social network or a news media? In: **Proceedings of the 19th international conference on World wide web**. [S.l.: s.n.], 2010. p. 591–600. Citado na página 31.

LANDIM, P. M. B. Sobre geoestatística e mapas. **Terrae Didatica**, v. 2, n. 1, p. 19–33, 2006. Citado nas páginas 58, 59 e 61.

LI, H.; CARAGEA, D.; CARAGEA, C.; HERNDON, N. Disaster response aided by tweet classification with a domain adaptation approach. **Journal of Contingencies and Crisis Management**, Wiley Online Library, v. 26, n. 1, p. 16–27, 2018. Citado na página 99.

LIU, Y.; JIANG, C.; ZHAO, H. Using contextual features and multi-view ensemble learning in product defect identification from online discussion forums. **Decision Support Systems**, Elsevier, v. 105, p. 1–12, 2018. Citado na página 95.

LONGUEVILLE, B. D.; LURASCHI, G.; SMITS, P.; PEEDELL, S.; GROEVE, T. D. Citizens as sensors for natural hazards: A vgi integration workflow. **Geomatica**, v. 64, n. 1, p. 41–59, 2010. Citado na página 35.

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: **Proceedings of the 2009 international workshop on location based social networks**. [S.l.: s.n.], 2009. p. 73–80. Citado na página 31.

LOPES, B. L. **Detecção de cenas em segmentos semanticamente complexos**. Tese (Doutorado) — Universidade de São Paulo, 2015. Citado nas páginas 41, 42, 43, 44, 95 e 140.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007. Citado na página 53.

MAKARIEVA, A. M.; GORSHKOV, V. G.; SHEIL, D.; NOBRE, A. D.; BUNYARD, P.; LI, B.-L. Why does air passage over forest yield more rain? examining the coupling between rainfall, pressure, and atmospheric moisture content. **Journal of Hydrometeorology**, v. 15, n. 1, p. 411–426, 2014. Citado na página 86.

MARTINS, J. A. **Efeito dos núcleos de condensação na formação de nuvens e o desenvolvimento da precipitação na região amazônica durante a estação seca**. 170–197 p. Tese (Doutorado) — Universidade de São Paulo, 2006. Citado na página 86.

MATHERON, G. Principles of geostatistics. **Economic geology**, Society of Economic Geologists, v. 58, n. 8, p. 1246–1266, 1963. Citado na página 58.

MATSUBARA, E. T.; MARTINS, C. A.; MONARD, M. C. **Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words**. [S.l.], 2003. v. 209, n. 4. Citado na página 46.

MCKINNEY, W. **pandas: powerful Python data analysis toolkit**. 2020. Acessado 06 dez. 2020. Disponível em: <https://pandas.pydata.org/pandas-docs/stable/pandas.pdf>. Citado na página 91.

MEIJERING, E. A chronology of interpolation: from ancient astronomy to modern signal and image processing. **Proceedings of the IEEE**, IEEE, v. 90, n. 3, p. 319–342, 2002. Citado na página 93.

MELGANI, F.; BRUZZONE, L. Classification of hyperspectral remote sensing images with support vector machines. **IEEE Transactions on geoscience and remote sensing**, IEEE, v. 42, n. 8, p. 1778–1790, 2004. Citado na página 53.

MENDIONDO, E. Reducing vulnerability ti water-related disasters in urban areas of the humid tropics. **Integrated Urban Water Management Humid Tropics, Paris, France**, p. 109–127, 2010. Citado nas páginas 30 e 36.

METHERON, G. Principles of geostatistics, economic geology. **Economic Geology**, v. 58, n. 8, p. 1246–1266, 1963. Citado na página 124.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013. Citado nas páginas 47 e 48.

MITCHELL, T. **Machine Learning**. [S.l.]: McGraw Hill, 1997. v. 1. Citado na página 48.

MIZUTORI, M.; GUHA-SAPIR, D. **Economic Losses, Poverty and Disasters 1998–2017**. [S.l.]: United Nations Office for Disaster Risk Reduction, 2017. Citado na página 30.

MÜLLNER, D. Modern hierarchical, agglomerative clustering algorithms. **arXiv preprint arXiv:1109.2378**, 2011. Citado na página 56.

NASCIMENTO, R. F. F.; ALCÂNTARA, E.; KAMPEL, M.; STECH, J. L.; NOVO, E.; FONSECA, L. M. G. O algoritmo support vector machines (svm): avaliação da separação ótima de classes em imagens ccd-cbers-2. **Simpósio Brasileiro de Sensoriamento Remoto**, v. 14, p. 2079–2086, 2009. Citado na página 53.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAN, W. **Applied linear statistical models**. [S.l.]: Irwin Chicago, 1996. Citado na página 55.

NIGAY, L.; COUTAZ, J. A design space for multimodal systems: concurrent processing and data fusion. In: **Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems**. [S.l.: s.n.], 1993. p. 172–178. Citado na página 40.

NORRIS, F. H.; STEVENS, S. P.; PFEFFERBAUM, B.; WYCHE, K. F.; PFEFFERBAUM, R. L. Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. **American journal of community psychology**, Springer, v. 41, n. 1-2, p. 127–150, 2008. Citado nas páginas 30 e 36.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP São Paulo, 2004. Citado na página 54.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado nas páginas 50, 101, 103 e 104.

PORIA, S.; CAMBRIA, E.; HOWARD, N.; HUANG, G.-B.; HUSSAIN, A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. **Neurocomputing**, Elsevier, v. 174, p. 50–59, 2016. Citado nas páginas 95 e 119.

POSER, K.; DRANSCH, D. Volunteered geographic information for disaster management with application to rapid flood damage estimation. **Geomatica**, v. 64, n. 1, p. 89–98, 2010. Citado nas páginas 30, 36, 37, 74 e 81.

POUYANFAR, S.; TAO, Y.; TIAN, H.; CHEN, S.-C.; SHYU, M.-L. Multimodal deep learning based on multiple correspondence analysis for disaster management. **World Wide Web**, Springer, v. 22, n. 5, p. 1893–1911, 2019. Citado nas páginas 93 e 119.

RESTREPO-ESTRADA, C.; ANDRADE, S. C. de; ABE, N.; FAVA, M. C.; MENDIONDO, E. M.; ALBUQUERQUE, J. P. de. Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring. **Computers & Geosciences**, Elsevier, v. 111, p. 148–158, 2018. Citado na página 82.

ROSA, K. D.; SHAH, R.; LIN, B.; GERSHMAN, A.; FREDERKING, R. Topical clustering of tweets. **Proceedings of the ACM SIGIR: SWSM**, v. 63, 2011. Citado na página 119.

ROSSI, A. L. D. **Ajuste de parâmetros de técnicas de classificação por algoritmos bioinspirados**. Tese (Doutorado) — Universidade de São Paulo, 2009. Citado na página 98.

ROY, J.; WARK, S. **Concepts, models, and tools for information fusion**. [S.l.]: Artech House, 2007. Citado na página 37.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: ACM. **Proceedings of the 19th international conference on World wide web**. [S.l.], 2010. p. 851–860. Citado nas páginas 31, 50, 63, 64, 73, 74 e 77.

SALAS, A.; GEORGAKIS, P.; PETALAS, Y. Incident detection using data from social media. In: IEEE. **2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)**. [S.l.], 2017. p. 751–755. Citado nas páginas 45, 50, 119 e 121.

SANH, V.; DEBUT, L.; CHAUMOND, J.; WOLF, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. **arXiv preprint arXiv:1910.01108**, 2019. Citado na página 146.

SCIKIT-LEARN. **sklearn.cluster.AgglomerativeClustering**. 2020. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>. Acesso em: 15/11/2020. Citado nas páginas 56 e 57.

____. **sklearn.cluster.OPTICS**. 2020. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html>. Acesso em: 15/11/2020. Citado na página 56.

____. **2.3. Clustering**. Scikit-Learn, 2021. Acessado em: 18/02/2021. Disponível em: <https://scikit-learn.org/stable/modules/clustering.html>. Citado na página 120.

____. **User Guide**. 2021. Disponível em: <https://scikit-learn.org/stable/user_guide.html>. Acesso em: 20/06/2021. Citado na página 101.

SHRYOCK, H. S.; SIEGEL, J. S.; LARMON, E. A. **The methods and materials of demography**. [S.l.]: US Bureau of the Census, 1973. v. 2. Citado na página 93.

SHYU, M.-L.; SARINNAPAKORN, K.; KURUPPU-APPUHAMILAGE, I.; CHEN, S.-C.; CHANG, L.; GOLDRING, T. Handling nominal features in anomaly intrusion detection problems. In: IEEE. **15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications (RIDE-SDMA'05)**. [S.l.], 2005. p. 55–62. Citado na página 93.

SIMON, P. **Too big to ignore: the business case for big data**. [S.l.]: John Wiley & Sons, 2013. v. 72. Citado na página 48.

SINOARA, R. A. **Aspectos semânticos na representação de textos para classificação automática**. Tese (Doutorado) — Universidade de São Paulo, 2018. Citado nas páginas 44, 45, 46 e 47.

SINOARA, R. A.; ANTUNES, J.; REZENDE, S. O. Text mining and semantics: a systematic mapping study. **Journal of the Brazilian Computer Society**, Springer, v. 23, n. 1, p. 9, 2017. Citado na página 47.

SPARKS, K.; THAKUR, G.; PASARKAR, A.; URBAN, M. A global analysis of cities' geosocial temporal signatures for points of interest hours of operation. **International Journal of Geographical Information Science**, Taylor & Francis, v. 34, n. 4, p. 759–776, 2020. Citado nas páginas 119 e 120.

SPINSANTI, L.; OSTERMANN, F. Automated geographic context analysis for volunteered information. **Applied Geography**, Elsevier, v. 43, p. 36–44, 2013. Citado na página 31.

STEIGER, E.; ALBUQUERQUE, J. P. D.; ZIPF, A. An advanced systematic literature review on spatiotemporal analyses of twitter data. **Transactions in GIS**, Wiley Online Library, v. 19, n. 6, p. 809–834, 2015. Citado na página 30.

SUNDERMANN, C.; ANTUNES, J.; DOMINGUES, M.; REZENDE, S. Exploration of word embedding model to improve context-aware recommender systems. In: IEEE. **2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)**. [S.l.], 2018. p. 383–388. Citado na página 47.

TALAIA, M.; VIGÁRIO, C. Temperatura de ponto de orvalho: um risco ou uma necessidade. **Geografia, cultura e riscos: livro de homenagem ao Prof. Doutor António Pedrosa**, Imprensa da Universidade de Coimbra, 2016. Citado na página 86.

TOBLER, W. R. A computer movie simulating urban growth in the detroit region. **Economic geography**, Taylor & Francis, v. 46, n. sup1, p. 234–240, 1970. Citado nas páginas 32, 91 e 121.

TUCCI, C. E.; BERTONI, J. C. *et al.* **Inundações urbanas na América do Sul**. [S.l.]: Ed. dos Autores, 2003. Citado nas páginas 29 e 86.

UNISDR. **2009 UNISDR Terminology on Disaster Risk Reduction**. [S.l.]: United Nations International Strategy for Disaster Reduction, 2009. Citado na página 35.

VIEIRA, S. R. *et al.* Geoestatística em estudos de variabilidade espacial do solo. **Tópicos em ciência do solo. Viçosa: Sociedade Brasileira de Ciência do Solo**, v. 1, p. 1–53, 2000. Citado na página 58.

VIVACQUA, A. S.; BORGES, M. R. Taking advantage of collective knowledge in emergency response systems. **Journal of Network and Computer Applications**, Elsevier, v. 35, n. 1, p. 189–198, 2012. Citado nas páginas 36 e 37.

WIN, S. S. M.; AUNG, T. N. Target oriented tweets monitoring system during natural disasters. In: IEEE. **2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)**. [S.l.], 2017. p. 143–148. Citado na página 30.

WINARNO, E.; HADIKURNIAWATI, W.; ROSSO, R. N. Location based service for presence system using haversine method. In: IEEE. **2017 International Conference on Innovative and Creative Information Technology (ICITech)**. [S.l.], 2017. p. 1–4. Citado na página 94.

XIE, Z.; GUAN, L. Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. In: IEEE. **2013 IEEE International Conference on Multimedia and Expo (ICME)**. [S.l.], 2013. p. 1–6. Citado nas páginas 32 e 40.

XU, C.-D.; WANG, J.-F.; HU, M.-G.; LI, Q.-X. Interpolation of missing temperature data at meteorological stations using p-bshade. **Journal of Climate**, v. 26, n. 19, p. 7452–7463, 2013. Citado na página 93.

YANG, Y.; POUYANFAR, S.; TIAN, H.; CHEN, M.; CHEN, S.-C.; SHYU, M.-L. If-mca: Importance factor-based multiple correspondence analysis for multimedia data analytics. **IEEE Transactions on Multimedia**, IEEE, v. 20, n. 4, p. 1024–1032, 2017. Citado na página 93.

YERVA, S. R.; JEUNG, H.; ABERER, K. Cloud based social and sensor data fusion. In: IEEE. **2012 15th International Conference on Information Fusion**. [S.l.], 2012. p. 2494–2501. Citado na página 41.

YIN, H.; LI, C. Human impact on floods and flood disasters on the yangtze river. **Geomorphology**, Elsevier, v. 41, n. 2-3, p. 105–109, 2001. Citado nas páginas 124 e 144.

YIN, J.; LAMPERT, A.; CAMERON, M.; ROBINSON, B.; POWER, R. Using social media to enhance emergency situation awareness. **IEEE intelligent systems**, IEEE, n. 6, p. 52–59, 2012. Citado nas páginas 31, 50, 64, 65, 72, 74 e 99.

ZADEH, A.; CHEN, M.; PORIA, S.; CAMBRIA, E.; MORENCY, L.-P. Tensor fusion network for multimodal sentiment analysis. **arXiv preprint arXiv:1707.07250**, 2017. Citado na página 95.

ZHANG, W.; YOSHIDA, T.; TANG, X. A comparative study of tf* idf, lsi and multi-words for text classification. **Expert Systems with Applications**, Elsevier, v. 38, n. 3, p. 2758–2765, 2011. Citado na página 46.

APÊNDICE

A

DICIONÁRIO DE *HASHTAGS*

A seguir na Tabela 31 e Tabela 32, observam-se as listas de *hashtags* com seus respectivos significados de estrangeirismos e palavras informais. Inclusive, essas informações foram captadas dos *tweets* do município de São Paulo do período de Novembro de 2016 até Outubro de 2018.

Além disso, observa-se na Tabela 31 as 50 *hashtags* mais frequentes entre as mensagens publicadas no Twitter relacionadas com alagamentos, por exemplo: “#chuva”, “#alagamento”, entre outras, inclusive, nota-se também a presença da frequência das *hashtags* nas mensagens publicadas no Twitter e a respectiva tradução para a norma culta da Língua Portuguesa.

Por último na Tabela 32, notam-se as 50 *hashtags* mais frequentes entre os *tweets* não relacionados com inundações, por exemplo: “#fotografia”, “#correndo”, “#bomdia”, entre outras, aliás observa-se também a presença da quantidade de ocorrências das *hashtags* nos *tweets* e a respectiva tradução para a norma culta da Língua Portuguesa.

Tabela 31 – Tradução das *hashtags* relacionadas com fenômenos naturais

hashtag	**quantidade de ocorrências**	**tradução**
#chuva	1168	Chuva
#cidadedagaroa	538	Cidade da garoa
#terradagaroa	424	Terra da garoa
#rain	385	Chuva
#spdagaroa	381	São Paulo cidade da garoa
#frio	97	Frio
#garoa	69	Garoa
#cantandonachuva	50	Cantando na chuva
#saopaulodagaroa	45	São Paulo cidade da garoa
#chuvinha	37	Chuva
#temporal	34	Temporal

#chovechuva	32	Caia chuva
#diadechuva	25	Dia de chuva
#apropositodachuva	25	A propósito da chuva
#guardachuva	23	Guarda-chuva
#diachuvoso	23	Dia de muita chuva
#chuvaboa	20	Ótima chuva
#chuvaemsp	19	Chuva em São Paulo
#cidadesagaroa	16	Cidade da garoa
#chuvadeverao	14	Chuva de verão
#chuvaesol	13	Chuva e sol
#chuvasp	13	Chuva em São Paulo
#chuvinhaboa	12	Ótima chuva
#alagamento	12	Alagamento
#spdagaroa463	11	São Paulo cidade da garoa
#diluvio	11	Dilúvio
#terradagaroasp	11	São Paulo cidade da garoa
#choveu	10	Choveu
#vemchuva	10	Vem chuva
#sãopaulodagaroa	10	São Paulo cidade da garoa
#depoisdachuva	9	Depois da chuva
#chuvoso	9	Muita chuva
#chovendo	9	Chovendo
#chove	9	Chove
#vaichover	8	Vai chover
#spterradagaroa	8	São Paulo cidade da garoa
#chuvisco	8	Chuvisco
#tachovendo	8	Está chovendo
#domingodechuva	7	Domingo de chuva
#capadechuva	7	Capa de chuva
#solechuva	7	Sol e chuva
#chuvaemsampa	6	Chuva em São Paulo
#chuvaforte	5	Chuva forte
#chuvaefrio	5	Chuva e frio
#nosolnachuva	5	No sol e na chuva
#corridanachuva	5	Corrida na chuva
#muitachuva	5	Muita chuva
#choveemsp	5	Chove em São Paulo

#antesdachuva	5	Antes da chuva
#enchente	5	Enchente

Fonte: Elaborada pelo autor.

Tabela 32 – Tradução das *hashtags* não relacionadas com fenômenos naturais

hashtag	**quantidade de ocorrências**	**tradução**
#saopaulo	858	São Paulo
#sp	828	São Paulo
#sampa	508	São Paulo
#splovers	293	Amando em São Paulo
#saopaulocity	249	Cidade de São Paulo
#saopaulowalk	225	Caminhando em São Paulo
#bomdia	147	Bom dia
#sãopaulo	136	São Paulo
#sousampa	129	Sou de São Paulo
#olharesdesampa	118	Olhares de São Paulo
#ig_saopaulo	111	São Paulo
#brasil	109	Brasil
#sp4you	101	São Paulo por você
#011	86	São Paulo
#sampacity	83	Cidade de São Paulo
#boanoite	73	Boa noite
#tbt	72	Quinta-feira passada
#avenidapaulista	64	Avenida Paulista
#boatarde	64	Boa tarde
#repost	63	Publicando novamente
#domingo	62	Domingo
#brazil	62	Brasil
#tvminuto	60	Minuto de televisão
#nofilter	59	sem filtro
#paulista	56	Paulista
#carnaval	54	Carnaval
#santos	54	Santos
#essepe	54	São Paulo
#sky	53	Céu
#samsung	52	Samsung
#sol	52	Sol

#photography	52	Fotografia
#spcity	50	Cidade de São Paulo
#city	48	Cidade
#céu	45	Céu
#rainyday	44	Dia de chuva
#love	44	Amor
#nikon	42	Fotografia
#storm	37	Tempestade
#reisen	36	Viagem
#sunset	36	Pôr do sol
#streetmagazine	36	Revista de rua
#photo	36	Fotografia
#tempestade	35	Tempestade
#arcoiris	34	Arco-íris
#presente	32	Presente
#sdv	32	Segue de volta
#instagood	32	Orgulho desta fotografia
#running	30	Correndo
#sabado	29	Sábado

Fonte: Elaborada pelo autor.

APÊNDICE

B

DICIONÁRIO DE PALAVRAS-CHAVE COLOQUIAIS

A seguir na Tabela 33, observam-se algumas das palavras-chave coloquiais mais frequentes presentes nas mensagens publicadas no Twitter e suas respectivas traduções para a norma culta da Língua Portuguesa. Inclusive, as traduções dessas palavras-chave informais são empregadas no mecanismo computacional capaz de efetuar a etapa de pré-processamento da metodologia de MT da abordagem de Fusão Multimodal e do *software* de identificação de possíveis vítimas de alagamentos.

Tabela 33 – Conversão de palavras-chave informais para formais

Informal	**Formal**
spdagaroa	São Paulo cidade da garoa
garôa	garoa
garoinha	garoa
chuvinha	chuva
tchuvinha	chuva
diaaa	bom dia
tardeeee	boa tarde
mobgraphia	evento cultural
mobgrafia	evento cultural
aniver	aniversário
sorvetineo	sorvete
trampar	trabalhar
trampando	trabalhando
turistar	visitar
turistando	visitando

chuvão	chuva
chóvens	jovens
buaaa	chorar
sabadinho	sábado
éssipe	São Paulo
heee	risada
unfollow	parar de seguir
splovers	admiradores de São Paulo
seloco	você é louco
turistada	visitar
cooore	correr
pakas	muito
busao	ônibus
busão	ônibus
orbourne	osbourne

Fonte: Elaborada pelo autor.

APÊNDICE

C

DICIONÁRIO DE PALAVRAS-CHAVE INFORMAIS DE OCORRÊNCIAS DE ALAGAMENTOS

A seguir na Tabela 34, observam-se algumas das abreviações e palavras-chave coloquiais mais frequentes encontradas nas ocorrências históricas de alagamentos da cidade de São Paulo reportadas pelo CGE-SP e suas respectivas traduções para a norma culta da Língua Portuguesa. Inclusive, essas traduções são utilizadas como pré-requisito para a execução do mecanismo computacional responsável pelo pré-processamento dos dados históricos de alagamentos.

Tabela 34 – Conversão de palavras-chave coloquiais e abreviações para a norma culta da Língua Portuguesa

Formal	**Informal**
AV.	AVENIDA
AVEN.	AVENIDA
PTE.	PONTE
PT.	PONTE
R.	RUA
PC.	PRAÇA
TN.	TÚNEL
JORN.	JORNALISTA
JORNAL.	JORNALISTA
PROF.	PROFESSOR
ES.	ESTRADA
EST.	ESTRADA
LG.	LARGO

VD.	VIADUTO
VELHA FEPASA	COMUNIDADE HUNGARA

Fonte: Elaborada pelo autor.

APÊNDICE

D

DICIONÁRIO DE EXPRESSÕES COLOQUIAIS DE OCORRÊNCIAS DE ALAGAMENTOS

A seguir na Tabela 35, notam-se algumas expressões coloquiais pertencentes as ocorrências de alagamentos do portal *Web* do CGE-SP que são suscetíveis ao processo de remoção. Aliás, essas expressões são utilizadas como pré-requisito para a execução do algoritmo responsável pelo pré-processamento dos dados históricos de alagamentos.

Tabela 35 – Remoções de expressões coloquiais

Expressões coloquias
ALTURA DO NÚMERO
ALTURA DO N.
ALTURA DO Nº
ALTURA DO NUMERO
ALTURA DO NUMERO.
ALT. Nº
ALT Nº
ALT. NÚMERO
ALT. DO N.
ALT. DO Nº
ALT. NR
ALT. N.
ALT N
ALT
ALTURA
ACESSO
TODA EXTENSÃO

SOB
PROX
INICIO DO MESMO
NO MESMO
MEIO DO MESMO
ENTRE
ANTES DO DESEMBOQUE
M ANTES
M APÓS
M ANTES DO DESEMBOQUE
M ANTES DA MESMA
METROS ANTES DA
. ANTES
. APÓS

Fonte: Elaborada pelo autor.

APÊNDICE

E

RESULTADOS DO PROCESSO DE TREINAMENTO DO MODELO DE FUSÃO MULTIMODAL DE MODO PRÉVIO

A seguir na Tabela 36 observam-se os resultados dos 50 testes realizados com os mecanismos de classificação utilizados no modelo de Fusão Multimodal de modo prévio. Aliás, os resultados estão organizados em ordem decrescente da métrica de avaliação chamada precisão.

Tabela 36 – Todos os resultados do processo de treinamento dos algoritmos de classificação do modelo de Fusão Multimodal de modo prévio

Classi-ficação	**Algoritmo**	**Estratégia de transformação de dados textuais para numéricos**	**Precisão**	**Recall**	**F1 -Score**
1	RF	BOW	0,8715	0,8667	0,8691
2	DT	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,8702	0,8444	0,8550
3	NB	BOW	0,8659	0,8659	0,8659
4	DT	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8635	0,8278	0,8465
5	RF	TF-IDF	0,8626	0,8547	0,8562
6	DT	TF-IDF	0,8605	0,8381	0,8495
7	DT	BOW	0,8594	0,8278	0,8433
8	DT	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,8591	0,8222	0,8427

9	DT	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,8561	0,8222	0,8396
10	RF	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,8541	0,8500	0,8516
11	RF	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,8519	0,8492	0,8509
12	LR	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,8516	0,8444	0,8494
13	DT	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,8511	0,8278	0,8392
14	RF	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,8483	0,8380	0,8431
15	RF	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8465	0,8436	0,8450
16	RF	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,8451	0,8380	0,8443
17	NB	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,8450	0,8202	0,8286
18	DT	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,8448	0,8167	0,8272
19	RF	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,8444	0,8333	0,8384
20	RF	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,8402	0,8389	0,8395

21	RF	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,8399	0,8333	0,8361
22	NB	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,8363	0,8222	0,8234
23	NB	TF-IDF	0,8362	0,8325	0,8338
24	DT	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,8360	0,8212	0,8286
25	NB	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,8360	0,8222	0,8314
26	DT	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,8317	0,8000	0,8172
27	LR	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,8312	0,8167	0,8228
28	LR	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8310	0,8090	0,8176
29	LR	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,8307	0,8045	0,8174
30	NB	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,8305	0,8167	0,8259
31	LR	BOW	0,8302	0,8167	0,8238
32	NB	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,8290	0,8111	0,8200
33	LR	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,8289	0,8212	0,8251
34	LR	TF-IDF	0,8278	0,8111	0,8205

35	NB	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8261	0,8100	0,8186
36	LR	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,8250	0,8100	0,8180
37	LR	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,8221	0,8044	0,8100
38	LR	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,8152	0,8056	0,8098
39	NB	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,8147	0,8111	0,8129
40	NB	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,7670	0,7487	0,7577
41	SVM	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,6954	0,6222	0,6527
42	SVM	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,6865	0,6111	0,6422
43	SVM	BOW	0,6801	0,6056	0,6409
44	SVM	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,6737	0,5889	0,6316
45	SVM	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,6680	0,6056	0,6407
46	SVM	TF-IDF	0,6678	0,6111	0,6495
47	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,6646	0,6145	0,6424
48	SVM	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,6577	0,5944	0,6298

49	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,6544	0,5978	0,6384
50	SVM	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,6531	0,6111	0,6327

Fonte: Elaborada pelo autor.

APÊNDICE

F

RESULTADOS DO PROCESSO DE TREINAMENTO DA ABORDAGEM DE IDENTIFICAÇÃO DE *TWEETS* RELACIONADOS COM ALAGAMENTOS DO MODELO DE FUSÃO MULTIMODAL DE MODO TARDIO

A seguir na Tabela 37 observam-se os resultados dos 50 testes realizados com os mecanismos de classificação utilizados na abordagem de identificação de *tweets* relacionados com alagamentos do modelo de Fusão Multimodal de modo tardio. Aliás, os resultados estão organizados em ordem decrescente da métrica de avaliação chamada precisão.

Tabela 37 – Todos resultados do processo de treinamento dos algoritmos de classificação da abordagem de identificação de *tweets* relacionados com enchentes do modelo de Fusão Multimodal de modo tardio

Classi-ficação	**Algoritmo**	**Estratégia de transformação de dados textuais para numéricos**	**Precisão**	**Recall**	**F1 -Score**
1	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,7950	0,7874	0,7912
2	SVM	BOW	0,7915	0,7885	0,7887
3	SVM	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,7873	0,7715	0,7786

4	SVM	TF-IDF	0,7855	0,7816	0,7836
5	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,7839	0,7701	0,7769
6	RF	BOW	0,7819	0,7772	0,7786
7	SVM	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,7811	0,7759	0,7780
8	LR	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,7763	0,7715	0,7736
9	LR	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,7752	0,7759	0,7755
10	SVM	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,7751	0,7644	0,7697
11	LR	BOW	0,7725	0,7644	0,7684
12	LR	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,7718	0,7701	0,7709
13	RF	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,7707	0,7701	0,7700
14	RF	TF-IDF	0,7706	0,7644	0,7675
15	LR	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,7703	0,7701	0,7702
16	LR	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,7703	0,7599	0,7650
17	RF	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,7697	0,7701	0,7699
18	SVM	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,7696	0,7644	0,7670

19	SVM	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,7694	0,7657	0,7675
20	LR	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,7694	0,7657	0,7676
21	SVM	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,7631	0,7429	0,7535
22	LR	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,7628	0,7542	0,7584
23	LR	TF-IDF	0,7600	0,7586	0,7593
24	NB	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,7595	0,7471	0,7510
25	DT	BOW	0,7593	0,7543	0,7568
26	RF	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,7565	0,7543	0,7554
27	NB	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,7548	0,7542	0,7545
28	RF	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,7537	0,7529	0,7533
29	RF	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,7536	0,7529	0,7532
30	RF	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,7532	0,7529	0,7530
31	RF	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,7489	0,7429	0,7440
32	NB	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,7448	0,7299	0,7381

33	NB	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,7410	0,7356	0,7383
34	LR	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,7409	0,7371	0,7390
35	DT	TF-IDF	0,7380	0,7257	0,7318
36	RF	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,7363	0,7314	0,7338
37	NB	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,7295	0,7126	0,7210
38	NB	TF-IDF	0,7272	0,7241	0,7257
39	NB	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,7237	0,7069	0,7122
40	NB	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,7224	0,7069	0,7145
41	NB	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,7195	0,7126	0,7172
42	NB	BOW	0,7142	0,6954	0,7105
43	DT	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,6810	0,6782	0,6796
44	DT	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,6760	0,6667	0,6713
45	DT	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,6743	0,6724	0,6734
46	DT	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,6568	0,6552	0,6560

47	DT	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,6413	0,6400	0,6406
48	DT	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,6341	0,6229	0,6285
49	DT	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,6127	0,6092	0,6113
50	DT	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,6064	0,6057	0,6061

Fonte: Elaborada pelo autor.

APÊNDICE

G

RESULTADOS DO PROCESSO DE TREINAMENTO DO MODELO DE FUSÃO MULTIMODAL DE MODO HÍBRIDO COM FOCO NA DECISÃO PROPORCIONADA PELOS DADOS METEOROLÓGICOS

A seguir na Tabela 38 observam-se os resultados dos 50 testes realizados com os mecanismos de classificação utilizados na abordagem de combinação de informações multimodais com foco na decisão proporcionada pelas informações meteorológicas. Aliás, os resultados estão organizados em ordem decrescente da métrica de avaliação chamada precisão.

Tabela 38 – Todos resultados do processo de treinamento dos algoritmos de classificação da abordagem de combinação de dados multimodais com foco na decisão proporcionada pelos dados meteorológicos

Classificação	**Algoritmo**	**Estratégia de transformação de dados textuais para numéricos**	**Precisão**	**Recall**	**F1-Score**
1	DT	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,8721	0,8722	0,8722
2	DT	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,8721	0,8715	0,8718
3	RF	TF-IDF	0,8699	0,8652	0,8679

4	SVM	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,8684	0,8667	0,8675
5	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,8673	0,8611	0,8611
6	SVM	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,8636	0,8556	0,8600
7	DT	TF-IDF	0,8636	0,8556	0,8615
8	NB	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,8630	0,8604	0,8617
9	SVM	BOW	0,8623	0,8547	0,8585
10	LR	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,8619	0,8556	0,8598
11	RF	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,8616	0,8611	0,8613
12	SVM	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,8614	0,8611	0,8613
13	SVM	TF-IDF	0,8599	0,8556	0,8577
14	RF	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,8596	0,8596	0,8596
15	SVM	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,8585	0,8547	0,8566
16	DT	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,8573	0,8556	0,8564
17	RF	BOW	0,8569	0,8556	0,8562
18	DT	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,8562	0,8548	0,8555

19	DT	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,8560	0,8539	0,8550
20	DT	BOW	0,8556	0,8444	0,8506
21	LR	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8553	0,8491	0,8522
22	LR	BOW	0,8552	0,8500	0,8526
23	LR	TF-IDF	0,8545	0,8436	0,8490
24	LR	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,8543	0,8389	0,8465
25	RF	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,8543	0,8500	0,8521
26	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8539	0,8492	0,8515
27	LR	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,8535	0,8492	0,8517
28	RF	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,8529	0,8444	0,8486
29	RF	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,8525	0,8389	0,8478
30	RF	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8507	0,8493	0,8500
31	LR	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,8507	0,8492	0,8499
32	DT	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,8505	0,8500	0,8502

33	RF	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,8491	0,8444	0,8463
34	DT	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,8490	0,8444	0,8463
35	NB	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,8470	0,8444	0,8457
36	DT	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8468	0,8389	0,8428
37	RF	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,8467	0,8444	0,8456
38	LR	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,8443	0,8389	0,8413
39	NB	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,8337	0,8333	0,8335
40	NB	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,8328	0,8325	0,8327
41	NB	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,8294	0,8268	0,8281
42	NB	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,8221	0,8100	0,8165
43	LR	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,8216	0,8111	0,8163
44	NB	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,8056	0,8056	0,8056

45	LR	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,8047	0,7889	0,7967
46	SVM	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,8021	0,7889	0,7996
47	NB	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,7919	0,7889	0,7904
48	SVM	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,7583	0,7556	0,7569
49	NB	TF-IDF	0,7058	0,6815	0,6915
50	NB	BOW	0,6843	0,6667	0,6798

Fonte: Elaborada pelo autor.

APÊNDICE

H

RESULTADOS DO PROCESSO DE TREINAMENTO DO MODELO DE CLASSIFICAÇÃO TEXTUAL

A seguir na Tabela 39 observam-se os resultados dos 50 testes realizados com os mecanismos de classificação utilizados na abordagem de identificação de possíveis vítimas de alagamentos a partir de dados textuais. Aliás, os resultados estão organizados em ordem decrescente da métrica de avaliação chamada precisão.

Tabela 39 – Todos resultados do processo de treinamento dos algoritmos de classificação da abordagem de identificação de possíveis vítimas de alagamentos a partir de dados textuais

Classificação	**Algoritmo**	**Estratégia de transformação de dados textuais para numéricos**	**Precisão**	**Recall**	**F1-Score**
1	NB	BOW	0,7007	0,6222	0,6535
2	RF	TF-IDF	0,6867	0,6833	0,6867
3	SVM	TF-IDF	0,6788	0,6760	0,6774
4	RF	BOW	0,6721	0,6667	0,6693
5	NB	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,6709	0,6167	0,6334
6	NB	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,6644	0,5978	0,6136
7	NB	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,6630	0,5667	0,6066

8	NB	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,6624	0,6056	0,6384
9	RF	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,6614	0,6537	0,6584
10	LR	BOW	0,6599	0,6500	0,6520
11	NB	TF-IDF	0,6593	0,6556	0,6574
12	SVM	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,6579	0,6500	0,6539
13	LR	TF-IDF	0,6567	0,6500	0,6533
14	LR	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,6563	0,6258	0,6405
15	RF	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,6525	0,6389	0,6439
16	SVM	BOW	0,6498	0,6313	0,6427
17	NB	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,6470	0,6056	0,6201
18	RF	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,6464	0,6389	0,6426
19	RF	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,6462	0,6333	0,6397
20	RF	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,6439	0,6424	0,6432
21	NB	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,6428	0,5810	0,6116
22	LR	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,6400	0,6333	0,6366

23	LR	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,6380	0,6201	0,6270
24	RF	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,6369	0,6222	0,6252
25	RF	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,6359	0,6313	0,6335
26	SVM	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,6358	0,6278	0,6317
27	LR	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,6352	0,6257	0,6304
28	DT	TF-IDF	0,6343	0,6278	0,6310
29	NB	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,6334	0,5667	0,5871
30	SVM	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,6330	0,6202	0,6266
31	SVM	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,6309	0,6313	0,6311
32	NB	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,6305	0,5978	0,6114
33	SVM	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,6296	0,6257	0,6276
34	RF	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,6271	0,6258	0,6264
35	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,6267	0,6200	0,6234

36	LR	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,6264	0,6145	0,6204
37	LR	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,6264	0,6222	0,6241
38	SVM	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,6263	0,6201	0,6231
39	LR	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,6250	0,6200	0,6225
40	SVM	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,6186	0,6089	0,6120
41	DT	BOW	0,6168	0,6167	0,6167
42	LR	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,6116	0,6034	0,6053
43	DT	Word Embeddings do tipo FastText da categoria CBOW com 50 dimensões	0,5902	0,5866	0,5884
44	DT	Word Embeddings do tipo Word2Vec da categoria CBOW com 50 dimensões	0,5860	0,5611	0,5717
45	DT	Word Embeddings do tipo Word2Vec da categoria CBOW com 100 dimensões	0,5849	0,5810	0,5834
46	DT	Word Embeddings do tipo FastText da categoria Skip-gram com 100 dimensões	0,5763	0,5697	0,5730
47	DT	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 50 dimensões	0,5650	0,5644	0,5639
48	DT	Word Embeddings do tipo Word2Vec da categoria Skip-gram com 100 dimensões	0,5619	0,5611	0,5615

49	DT	Word Embeddings do tipo FastText da categoria Skip-gram com 50 dimensões	0,5424	0,5389	0,5406
50	DT	Word Embeddings do tipo FastText da categoria CBOW com 100 dimensões	0,5263	0,5222	0,5243

Fonte: Elaborada pelo autor.


ICMC
USP
SÃO CARLOS