UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE COMPUTAÇÃO

Proposta de Dissertação de Mestrado

Análise de Relevância de Dados Complexos para Detecção de Eventos

Caroline Mazini Rodrigues (Aluno)
Prof. Dr. Zanoni Dias (Orientador)
Prof. Dr. Anderson Rocha (Coorientador)

Resumo

Considerando a ocorrência de um evento com grande impacto social, é importante estabelecer uma relação espaço-temporal das informações disponíveis e assim, responder algumas questões sobre o evento como *"quem"*, *"como"*, *"onde"* e o *"porquê"*. Este trabalho é parte do projeto temático FAPESP *"DéjàVu: Feature-Space-Time Coherence from Heterogeneous Data for Media Integrity Analytics and Interpretation of Events"* e propõe, a partir de dados coletados de redes sociais, determinar a relevância destes para o evento analisado, possibilitando a correta construção de relações entre esses dados durante a etapa de análise. Os principais desafios deste trabalho são as características dos dados que serão utilizados: heterogeneidade, já que são obtidos de diferentes fontes; multimodalidade, incluindo textos, imagens e vídeos; não-rotulação, não apresentando rótulos de relevância para o evento; e não-estruturação, já que não possuem características que possam ser utilizadas diretamente durante o aprendizado. Para determinar a relevância de itens analisados, será seguida uma sequência de etapas que incluem: preparação dos dados, onde eliminaremos redundâncias e rotularemos parte dos dados; engenharia de características, onde extrairemos características visuais e textuais; aprendizado de máquina, onde analisaremos técnicas de aprendizado não-supervisionado e semi-supervisionado; e, validação e análise dos resultados, onde avaliaremos as soluções obtidas.

UNIVERSITY OF CAMPINAS

INSTITUTE OF COMPUTING

Master Dissertation Proposal

Complex Data Relevance Analysis for Event Detection

Caroline Mazini Rodrigues (Candidate)
Prof. PhD. Zanoni Dias (Advisor)
Prof. PhD. Anderson Rocha (Co-advisor)

Abstract

Considering the occurrence of an event with high social impact, it is important to establish a space-time relation of available information and so, answer some questions about the event as *"who"*, *"how"*, *"where"* and *"why"*. This work is part of the thematic FAPESP project *"DéjàVu: Feature-Space-Time Coherence from Heterogeneous Data for Media Integrity Analytics and Interpretation of Events"* and it proposes, from social network collected data, to determine the relevance of them for the analyzed event, allowing the correct construction of relationships among these data during an analysis phase later on. The main challenges of this work are the characteristics of the data which will be used: heterogeneity, as they come from different sources; multi-modality, such as texts, images and videos; unlabeled data, as they do not present label of straightforward relevance for the event; and unstructured data, as they do not possess characteristics which could be used directly during the learning. To determine the relevance of analyzed items, will be followed a sequence of phases which include: data preparation, in which we will eliminate redundancies and label some data; features engineering, in which we will extract visual and textual features; machine learning, in which we will analyze unsupervised and semi-supervised learning techniques; and validation and results analysis, in which we will evaluate the obtained solutions.

1 Introdução

Atualmente, grande parte da comunicação entre pessoas é realizada por meio de redes sociais como *Facebook, Instagram* e *Twitter* e, consequentemente, uma enorme quantidade de dados é transmitida quase instantaneamente. Muitos desses dados contêm informações referentes às vidas de usuários, incluindo gostos, cotidiano e possíveis locais e eventos que frequentaram. Parte dessas informações pode ser utilizada para diferentes finalidades como fornecer recomendações de busca [2, 26], traçar perfis de grupos de usuários [10, 20], e até mesmo obter informações de determinados eventos que ocorreram na hora e no local em que o usuário estava [11]. No entanto, além dessas informações, que podem ser consideradas úteis para algumas aplicações, existe a presença de ruídos como informações falsas e dados manipulados e/ou replicados [1, 23]. Um dos grandes desafios que surge na necessidade de trabalhar com essa quantidade massiva de dados é determinar sua importância para cada situação.

O problema da determinação de importância dos dados ganha ainda mais destaque quando objetiva-se tomar decisões críticas a partir da interpretação desses dados obtidos de redes sociais. Por exemplo, uma sequência de imagens tiradas durante, imediatamente antes ou depois de um ataque terrorista, e publicadas no *Twitter* ou *Facebook*, podem auxiliar na reconstrução do evento e, possivelmente, prover informações que indiquem *"quem"*, *"como"*, *"onde"* e o *"porquê"* dos acontecimentos [3]. No entanto, essa reconstrução não é uma tarefa trivial, já que não basta recuperar todos os dados disponíveis que possam ter relação com um evento; é necessário que estes tenham conexões entre si e sejam relevantes, para possibilitar uma composição coerente e propícia à interpretação do evento.

Com relação à conexão entre esses dados, uma abordagem lógica e de fácil interpretação seria a temporal, ou seja, ordená-los de forma que seja possível construir uma linha do tempo descritiva do evento. Essa ordenação cronológica possibilitaria o estabelecimento de relações de causa e efeito, identificação de possíveis atores e determinação de focos de interesse, no sentido temporal e espacial, para análise do evento. Entretanto, não basta apenas ordenar os dados cronologicamente se estes não apresentarem uma relevância adequada para a maximização da qualidade de descrição do evento em questão. Para que a análise de um evento seja completa; é necessário que os dados brutos sejam filtrados de acordo com sua semelhança com o que se deseja recuperar. Assim, textos, imagens e/ou vídeos mais relevantes podem ser utilizados no processo de ordenação cronológica, agregando complexidade ao problema.

A complexidade embutida na determinação de relevância desses dados deve-se, principalmente, à natureza de cada um deles. Estes constituem um conjunto heterogêneo, no que diz respeito às fontes, formas de aquisição e armazenamento; multimodal, referindo-se aos tipos de dados que podem incluir texto, imagens e vídeos; não-estruturado, já que não apresentam um padrão de representação; e não-rotulado, considerando que estes dados não vêm classificados como relevantes

ou não.

Como foco deste projeto, iremos explorar soluções para lidar com os desafios desses dados, com o intuito de determinar a relevância de textos, imagens e vídeos com relação a eventos de interesse. Pretendemos desenvolver uma solução dinâmica e geral, de forma que seja possível abranger a dinamicidade da obtenção dos dados a partir de mídias sociais e internet, e apresentar resultados satisfatórios mesmo quando analisados diferentes eventos. Para atingir esse objetivo, seguiremos uma sequência de etapas, incluindo: Preparação dos Dados, onde eliminaremos redundâncias e rotularemos parte dos dados; Engenharia de Características, onde extrairemos características visuais e textuais para posterior exploração das técnicas de projeção, transformação e fusão; Aprendizado, onde analisaremos técnicas de aprendizado não-supervisionado e semi-supervisionado com base nas características extraídas; e, Validação e Análise dos Resultados, onde avaliaremos as soluções obtidas com o uso de bases de dados de eventos disponíveis.

1.1 Impacto e Contribuições Científicas

Ao aprendermos a trabalhar com um conjunto de dados multimodal, hetorogêneo, não-rotulado e não-estruturado, características que constituem a maior parte dos dados disponíveis na atualidade, ampliará as possibilidades do que se pode compreender a partir da análise de dados. O uso restrito de dados bem-condicionados descarta um número indescritível de possibilidades, já que esses são escassos e, muitas vezes, não refletem a realidade. Trabalhos que utilizam dados "brutos" começam a ter maior destaque, dada a natureza dinâmica dos problemas a serem atacados e do poder apresentado por soluções criadas com base neles.

No entanto, a dificuldade em trabalhar com tantos dados e de diferentes qualidades é determinante para o sucesso de qualquer aplicação. Dessa maneira, definir a relevância de um dado para uma aplicação reduz o espaço de interesse, diminuindo tempo e poder necessário para processamento, além de aumentar a acurácia dos resultados obtidos. É possível afirmar que, o desempenho de um modelo é diretamente proporcional à qualidade e ao poder de descrição de um conjunto de entrada. Considerando o contexto geral do projeto, que é a análise e interpretação de eventos com foco forense, percebemos nitidamente que, a aplicabilidade da solução é ampla, auxiliando na compreensão de detalhes dos incidentes; portanto, a acurácia é muito importante para o sucesso do projeto, evitando falhas de julgamento.

Com relação às contribuições científicas, destacamos: o trabalho com dados "brutos" e em grande quantidade, sendo um desafio atual com possibilidades de melhorias tanto em relação ao tempo e poder de processamento quanto à acurácia dos resultados; o desenvolvimento de técnicas que possibilitem o uso de múltiplas modalidades de dados além de, possivelmente, a comparação entre elas; e, a análise de técnicas de aprendizado não-supervisionado e semi-supervisionado, que tem recebido espaço devido à necessidade crescente de aprender modelos sem a disponibilidade de um conjunto considerável de amostras de treinamento. É importante apontar que, este trabalho não

apresenta precedentes literários, sendo uma solução nova para aumentar o poder da computação forense no cenário mundial.

1.2 Organização

Este trabalho está organizado da seguinte maneira: na Seção 2 descrevemos alguns trabalhos do estado da arte que utilizam metodologias similares às abordadas neste projeto; na Seção 3 discutimos os problemas e possíveis desafios encontrados durante o desenvolvimento do projeto; nas Seções 4 e 5 apresentamos, respectivamente, os objetivos e a metodologia que será empregada na execução; na Seção 7 apresentamos e discutimos os resultados preliminares; por último, na Seção 6 detalhamos o plano de trabalho.

2 Estado da Arte

Com o intuito de lidar com a multimodalidade, podemos destacar três abordagens de agrupamento: projeção de características latentes em um plano comum entre as diferentes modalidades; descrição dos dados por meio de uma única modalidade; e tratamento individual de cada uma das modalidades seguida por combinação dos resultados. Neste trabalho, essas três abordagens serão chamadas de: projeção, transformação e fusão, respectivamente.

2.1 Projeção

Dentre as metodologias que procuram projetar características em um plano comum para todas as modalidades está a *Collective Component Analysis (CoCA)*. A *CoCA* [24] é baseada no princípio da *Análise de Componente Coletivo* que objetiva superar problemas de heterogeneidade e multimodalidade dos dados. Em um primeiro momento, a CoCA propõe a definição de um espaço de projeção em comum para características de dados diferentes. Na sequência, verifica-se a existência de relacionamentos entre esses dados e procura-se aproximar características relacionadas nesse espaço de projeção, de acordo com grafos de relacionamento.

Também no contexto de projeção de características em um plano comum, Ngiam *et al.* [18] apresentaram um método que utiliza *Deep Autoencoders* para descrever o espaço compartilhado de características de vídeo e áudio, tendo como objetivo implementar um modelo consistente tanto na presença de ambas as modalidades quanto na ausência de uma delas. Os autores destacaram que a *Cross-domain Feature (CCA)*, que tem como objetivo encontrar transformações que maximizem as correlações entre os dados, produz os melhores resultados, mas não atinge a robustez do método utilizando *Deep Autoencoders*.

O trabalho apresentado por Wang *et al.* [28] propõe uma análise de relações intra e intermodais de forma a possibilitar o mapeamento de características multimodais em um espaço comum. Os autores exploraram uma abordagem não-supervisionada e uma supervisionada. A primeira, não-supervisionada, foi chamada de *Multi-Modal Stacked Auto-Encoders (MSAE)*, sendo uma pilha de *Auto-Encoders* para cada modalidade, onde o objetivo é obter características latentes que unifiquem

as modalidades, possibilitando uma comparação semântica intermodal sem que se perca as relações semânticas intramodais. Já a abordagem supervisionada contou com o uso de um modelo chamado *Multi-Modal Deep Neural Network (MDNN)* baseado em *Deep Convolutional Network (DCNN)* e *Neural Language Model (NLM)*. Da mesma maneira que os *Auto-Encoders*, o MDNN também utiliza funções de mapeamento para cada modalidade, mas dessa vez, de forma supervisionada, utilizando a rede *AlexNet* para imagens e um *Skip-Gram Model (SGM)* para textos e, a partir das características latentes obtidas, ambas as modalidades são treinadas conjuntamente de forma a mapear os relacionamentos intermodais. Em comparação realizada no trabalho, os autores concluiram que o método supervisionado apresenta melhor performance.

No trabalho de Zhu e Xie [37] é explorado o *embedding* de eventos, onde estes podem ser projetados em um espaço comum para comparação. Os dados utilizados são textos de relatórios policiais contendo hora, local e descrição do incidente. São utilizadas *Restricted Boltzmann Machines (RBMs)* para selecionar características textuais visíveis mais importantes (palavras-chave) que serão utilizadas para compor o *embedding*. A abordagem utilizada é não-supervisionada já que as *RBMs* não necessitam de rótulos, obtendo padrões de coocorrência das variáveis. Uma RBM é um modelo indireto estocástico que possui uma camada visível e uma escondida que seguem a função de energia definindo uma distribuição de probabilidade para os parâmetros de entrada.

Por fim, contando com grande destaque na projeção de características está o *StarSpace* apresentado por Wu *et al.* [29], pesquisadores do Facebook. O *StarSpace* é um modelo de *embedding* com múltiplas finalidades, dentre elas: classificação de texto, ranking e recomendação multimodal. O princípio por trás desse *embedding* é o aprendizado de entidades que podem ser de diferentes modalidades e são descritas como *sacolas de palavras* obtidas a partir de um dicionário de tamanho fixo. Após o aprendizado, os vetores de características finais para cada entidade podem ser comparados, independentemente da modalidade. A partir dos experimentos realizados incluindo classificação de texto, predição de links, recomendação de documentos, busca por artigos, correspondência e aprendizado de *enbeddings* de frases, foi possível observar que o *StarSpace* apresenta bons resultados quando comparado aos métodos existentes.

2.2 Transformação

Com relação à tentativa de descrição de dados de todas as modalidades em uma única, podemos citar trabalhos como o desenvolvido por Singh *et al.* [25], onde o processo de reconhecimento de situações ocorre com o uso de dados heterogêneos e multimodais a partir da obtenção de características chamadas de *STTPoints* incluindo informações sobre *"o quê", "quando"* e *"onde"* ocorreu a situação, seguindo por um processo de transformação dessas características em *E-mages* que são representações na forma de grades espaciais e, por fim, ocorrendo o processo de detecção de situações que, neste caso, é realizado por meio de classificação com base em conhecimentos de domínio. Seguindo a mesma abordagem, Yatskar *et al.* [33] representam uma imagem através dos

termos descritivos como ação principal, atores, objetos presentes na cena e local.

A ideia de descrição de modalidades em uma única é também explorada por Yang *et al.* [31], onde o objetivo é determinar o conceito contido em um vídeo utilizando um conjunto de imagens auxiliares com textos descritivos obtidos do *Flickr* e, utilizando a parte textual, extrair informações semânticas através do *Wikipedia, Microsoft N-gram Services* e do próprio *Flickr*. Além disso, a proposta do trabalho é ser capaz de utilizar apenas informações mais relevantes em seus classificadores de conceitos, ou seja, apenas os frames mais descritivos de um vídeo, já que alguns deles podem ser considerados ruído. Essa metodologia foi chamada de *Boosted Concept Learning*. Para determinar o que é ruído, o trabalho propõe ainda um método para aprendizado de características sensíveis da distribuição, que consiste em obter algumas entradas corrompidas ao, estocasticamente, atribuir valor 0 a alguns componentes e, assim, capturar dependências estatísticas.

2.3 Fusão

Por fim, abordagens que tratam individualmente cada uma das modalidades, em determinado estágio do processo de aprendizagem, buscam realizar a fusão das decisões em cada uma das modalidades. Um dos trabalhos que utiliza esse tipo de abordagem é apresentado por Srivastava e Salakhutdinov [27] que buscou trabalhar com as modalidades separadamente gerando características que pudessem ser consideradas livres de modalidade, utilizando uma *Deep Belief Network (DBN)*, que é um conjunto de RBMs, com duas camadas separadas, sendo que cada uma trata modalidades diferentes de dados, texto e imagem, e para compor o sistema multimodal foi utilizado uma RBM juntando as camadas especializadas.

De maneira semelhante, Feng *et al.* [8] tratam a multimodalidade de dados utilizando duas abordagens para comparação: uma pilha de *Correspondence Restricted Boltzmann Machine (Corr-RBM)* para cada modalidade, aprendendo relações de correspondência entre modalidades em cada camada do rede; e uma *Correspondence Deep Belief Network (Corr-DBN)*, para determinar a correspondência das modalidades na última camada, diminuindo as operações de correlação entre as modalidades. Os autores concluíram que a correlação nas primeiras camadas podem ser mais eficientes e mais fáceis de serem obtidas.

O modelo proposto por Qian *et al.* [21] objetiva utilizar dados multimodais de redes sociais para rastrear eventos e sua evolução e, posteriormente, resumir o contexto do evento com relação ao seu desenvolvimento temporal. Nesse trabalho, os autores dividem o conteúdo de mídia social em dois contextos, visual e não-visual. Como contexto visual, entende-se o conjunto de documentos que apresentam texto e imagens correspondentes. Já no contexto não-visual, é considerada a presença estrita de textos. Como observado pelos autores, as distribuições de características visuais e não-visuais são diferentes e, para abordar ambas, utilizaram o *Multi-Modal Event Topic Model (mmETM)* onde são criadas *Bags-of-Words* para descrever tanto imagens quanto textos. Em *mmETM*, os contextos visual e não-visual são tratados separadamente, gerando uma distribuição

para cada tópico de acordo com o contexto em que se encaixa, sendo que, para o contexto visual, a distribuição final é formada pela composição das distribuições de textos e imagens.

3 Análise e Interpretação de Eventos a partir de Dados Heterogêneos

Neste projeto de pesquisa, será feita a análise de relevância de dados relacionados a eventos coletados por meio de mídias sociais e da internet. Acredita-se que a ocorrência de determinados eventos em lugares públicos pode dispor de uma grande quantidade de informações obtidas por indivíduos que testemunharam os acontecimentos e as disponibilizaram em redes sociais. Saber utilizar esses dados pode proporcionar respostas a algumas questões forenses como: *"quem"*, *"como"*, *"onde"* e o *"porquê"*. No entanto, o uso de dados irrelevantes pode conduzir a falhas de julgamento ou impossibilidade em responder essas questões.

Em um contexto geral, essa pesquisa faz parte do projeto temático DéjàVu financiado pela *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP), descrito na Seção 3.1. Na Seção 3.2, apresentamos o foco principal deste trabalho, detalhando, nas seções seguintes, os principais problemas na extração de características (Seção 3.3) e na construção de um modelo de aprendizado de máquina (Seção 3.4) para tratá-lo.

3.1 Projeto DéjàVu

Dado um evento a ser analisado, é importante estabelecer uma relação espaço-temporal de forma a possibilitar a obtenção de informações significativas. O projeto temático FAPESP *"DéjàVu: Feature-Space-Time Coherence from Heterogeneous Data for Media Integrity Analytics and Interpretation of Events"*[1] propõe uma sequência de etapas para automatizar a análise de eventos sob investigação, contando com um grupo amplo de pesquisadores para este propósito.

Dentre as etapas desse processo de análise estão inclusas: coleta de dados de redes sociais, parceiros e/ou base de dados públicas (*Aquisição de Dados*); organização dos dados e estabelecimento da *Coerência-X* (*Organização dos Dados e Coerência-X*), onde *Coerência-X* denomina a sincronização dos dados de maneira coerente com características espaço-temporais, possibilitando a representação comum dos dados; e extração de informações e obtenção de significados (*Compreensão do Conteúdo e Inferência*). A Figura 1 mostra a sequência dessas etapas.

Cada uma dessas etapas pode ser descrita como um conjunto de atividades desenvolvidas durante o andamento do projeto:

Aquisição de Dados: nesta etapa, em um primeiro momento, os dados utilizados serão de bases de dados públicas possibilitando a avaliação dos algoritmos inicialmente desenvolvidos. Na sequência, serão agregados os dados obtidos por meio das parcerias como *National Institute of*

[1]Projeto Temático FAPESP 2017/12646-3, em execução de Dezembro de 2017 até Novembro de 2022 sob coordenação do Prof. Dr. Anderson Rocha

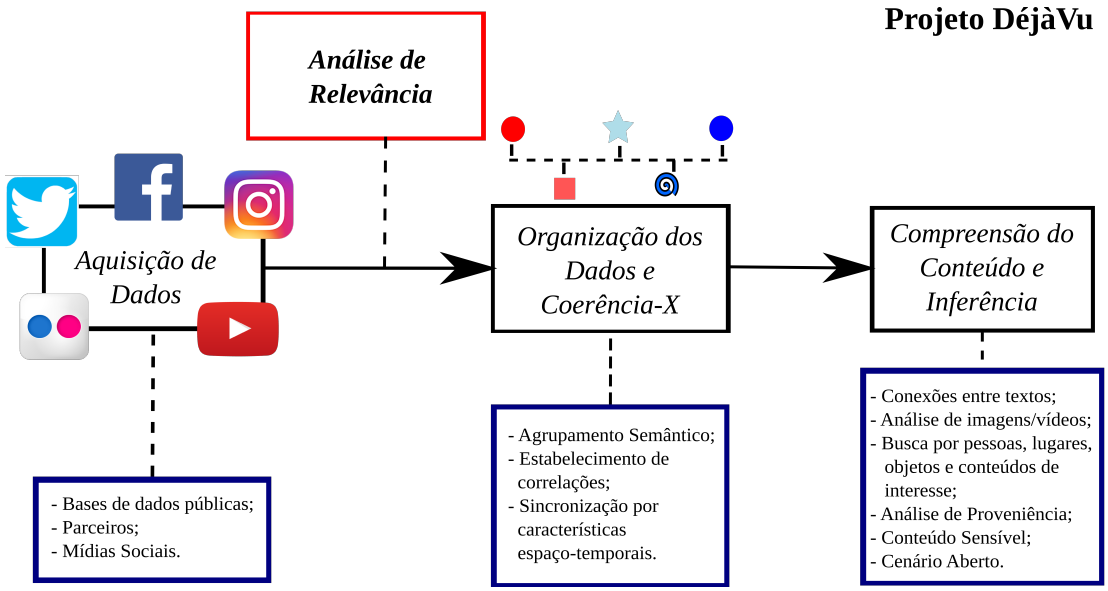

Figura 1: Etapas da análise de eventos do projeto DéjàVu. Inicialmente, os dados são coletados (*Aquisição de Dados*), passando por uma filtragem por relevância, seguindo para etapa de sincronização e correlação (*Organização dos Dados e Coerência-X*) e finalizando na etapa de extração de informações (*Compreensão do Conteúdo e Inferência*).

Standards and Technology (NIST), *Defense Advanced Research Projects Agency* (DARPA) e RankOne Inc.. Por fim, alguns eventos serão determinados para a coleta dos dados de mídias socias e aplicação da solução desenvolvida;

Organização dos Dados e Coerência-X: nesta fase do projeto, os dados coletados serão organizados de acordo com tópico e linha do tempo correspondentes, encontrando estruturas que melhor agrupem os dados semanticamente. Na sequência, serão determinados os grupos mais importantes e encontradas as correlações com o evento e, posteriormente, feita a associação das imagens com textos que possam apresentar significados semelhantes. Por fim, será explorada a *Coerência-X*;

Compreensão do Conteúdo e Inferência: nesta última etapa do projeto, será realizada a análise do conteúdo dos dados organizados buscando compreender o evento, condições e fatores envolvidos. Serão tratadas conexões entre textos, análise de imagens/vídeos em busca de pessoas, lugares, objetos e conteúdos de interesse e, definição de relacionamentos entre eles (análise de proveniência). Além disso, será desenvolvida a análise de mídias sensíveis, buscando determinar a existência de conteúdo que prejudique os usuários, como por exemplo, pornografia infantil e casos de violência. Como um desafio para esta etapa do projeto; destaca-se a necessidade de atuação em cenário aberto, ou seja, lidando com informações novas ao modelo de aprendizagem, de classes desconhecidas.

(a) Incêndio Grenfell. (b) Prédios não relacionados.

Figura 2: Exemplos de possíveis imagens retornadas para a busca *"Grenfell Tower Fire"*.

3.2 Análise de Relevância

A partir do momento em que um conjunto de dados é obtido e pretende-se estabelecer a *Coerência-X*, é importante determinar a relevância do que foi coletado para a análise que desejamos realizar. Por exemplo, escolhendo um evento específico como o incêndio da *Grenfell Tower*, que ocorreu em 14 de junho de 2017 em North Kensington, Londres, ocasionando 71 mortes e mais de 70 feridos, uma possível coleta de dados pode retornar qualquer dado relacionado com o prédio antes, durante ou depois do evento. Adicionalmente, até mesmo outros prédios ou assuntos que possam referenciar a *Grenfell Tower*, conforme exemplificado pela Figura 2, podem ser retornados. No entanto, o interesse é apenas na sucessão de fatos que ocorreram nas proximidades do evento, que se relacionem com este em um determinado período de tempo. Se uma filtragem dos dados de coleta não é realizada antes da construção da *Coerência-X*, possivelmente, as relações estabelecidas serão confusas e/ou pouco confiáveis, já que apresentam ruído e irrelevância.

O foco deste trabalho é a filtragem de dados de acordo com a relevância com um determinado evento, ou seja, análise de relevância (*Relevance Analysis*). Considerando um conjunto de dados extraídos de redes sociais, é importante determinar a existência de uma relação com o evento analisado. Para isso, é preciso superar alguns problemas, já mencionados anteriormente, que incluem heterogeneidade, multimodalidade, não-rotulação e não-estruturação.

Heterogeneidade: indica o nível de semelhança das fontes dos dados, formas de aquisição e armazenamento. Considerando que este projeto visa trabalhar com dados de redes sociais, dados fornecidos por parceiros e bases de dados públicas, espera-se alta heterogeneidade e portanto, será necesserário saber trabalhar em um espaço de características comum que possa abranger a grande maioria das fontes;

Multimodalidade: refere-se ao tipo do dado coletado, como por exemplo, textos, imagens ou vídeos. A importância em trabalhar com a multimodalidade deve-se à natureza semântica

complementar de cada modalidade. Em outras palavras, um texto pode apresentar palavras-chave que indiquem a ocorrência do evento buscado, uma imagem relacionada a esse texto provavelmente também está relacionada ao evento e pode fornecer informações sobre localidade e detalhes visuais do ocorrido, enquanto que um vídeo, além de agregar todas as informações contidas em imagens, pode ainda fornecer a percepção temporal;

Não-Rotulação: os rótulos neste caso indicariam a relevância ou não do dado de forma a ser possível realizar algum tipo de treinamento para a classificação de relevância. Todavia, como o objetivo principal do trabalho é encontrar dados em mídias sociais, não existe rotulação prévia, e uma rotulação posterior seria dificuldada pela mudança de eventos sob análise, prejudicando, inclusive, o tempo de resposta ao evento;

Não-Estruturação: considerando a heterogeneidade e a multimodalidade, os dados, na maioria das situações, não seguem um padrão de representação nem possuem características que possam ser utilizadas diretamente como critério de avaliação, dessa maneira, não estão estruturados.

3.3 Engenharia de Características

Como já mencionado anteriormente, a multimodalidade e não-estruturação promovem um desafio no quesito extração de características. Como trataremos de textos, imagens e vídeos, será necessário definir uma maneira de extrair essas características de dados não-estruturados de forma que as informações semânticas sejam mantidas.

Como previamente observado, três abordagens se destacam na literatura para este fim: projeção de características latentes em subespaço comum, descrição dos dados por meio de uma única modalidade e o tratamento individual de cada modalidade. Estas abordagens, nomeadas nesse trabalho como *Projeção, Tranformação* e *Fusão*, são descritas na sequência.

Projeção: consiste em encontrar um espaço comum onde as características de diferentes modalidades possam ser projetadas mantendo relacionamentos de vizinhança intra e entre classes. Projetar características multimodais em um plano comum facilitaria a posterior comparação de dados de modalidades diferentes; no entanto, essa projeção pode ocasionar perdas de informações importantes para o processo de análise do evento;

Transformação: consiste no uso de uma modalidade para representação de todos os demais dados, por exemplo, a obtenção da descrição de uma imagem transforma o conteúdo semântico visual em palavras representativas. Essa abordagem, no entanto, pode, assim como na projeção, apresentar perdas de informações para a modalidade a ser transformada, mesmo que o processo de comparação entre modalidades seja facilitado;

Fusão: consiste no tratamento de cada modalidade individualmente, sendo necessária a fusão dos resultados obtidos. Esta abordagem facilitaria a completa representação de cada uma das modalidades, entretanto, seria necessário determinar um método de fusão que possibilite a comparação entre elas.

3.4 Aprendizado de Máquina

Assim que as características são extraídas, é importante aprender um modelo que melhor separe os dados de acordo com sua relevância para o evento. Neste momento, surge o desafio dos dados não-rotulados, o que impossibilita o uso de técnicas de aprendizado supervisionado. Dessa maneira, explorar métodos de aprendizado não-supervisionado evitaria o problema da não-rotulação, no entanto, o desempenho dos métodos na ausência de qualquer conhecimento prévio pode ser prejudicado, assim, exploraremos também métodos de aprendizado semi-supervisionado para a determinação da relevância a partir das características extraídas.

Não-Supervisionado: neste tipo de aprendizado, os agrupamentos são formados de acordo com semelhanças obtidas a partir das características que representam cada amostra de dados. Algumas questões podem ser levantadas: número de grupos, medidas de similaridade e/ou medidas estatísticas a serem calculadas, como taxa de representatividade de cada grupo.

Semi-Supervisionado: neste tipo de aprendizado, é possível utilizar um conjunto de dados rotulados como sementes iniciais. Por outro lado, sendo este um conjunto pequeno, ocorre um espalhamento de rótulos por meio de métricas de proximidade. Algumas questões podem ser levantadas: escolha de dados mais representativos dos conjuntos para serem rotulados, métricas de distância e/ou medidas estatísticas a serem calculadas, além da forma de validar os rótulos atribuídos.

É importante destacar que, dentro da computação forense, o uso de intervenção humana é, em alguns casos, necessária, seja na rotulação de algumas amostras, no direcionamento do aprendizado ou na validação da metodologia, evitando erros de julgamento e adicionando maior confiabilidade ao modelo. Alguns exemplos de intervenção humana podem ser observados em alguns trabalhos relacionados à detecção de eventos sonoros [12] (rotulação) e ao reconhecimento de assinaturas [16] (validação).

4 Objetivos

No contexto deste trabalho; apresentamos um objetivo geral e um conjunto de objetivos específicos a serem alcançados.

4.1 Objetivo Geral

Como objetivo geral, pretendemos explorar soluções para a determinação de relevância de um dado quando considerado um evento específico, possibilitando o correto estabelecimento de coerência

espaço-temporal entre os dados sob análise na fase seguinte do projeto DéjàVu. É importante destacar que a fase de coerência espaço-temporal não está no escopo deste trabalho, mas utiliza os resultados obtidos aqui para o desenvolvimento da coerência. É esperado que as soluções finais sejam gerais, de forma a apresentar resultados satisfatórios quando analisados diferentes eventos; e dinâmicas, comportando o crescente aumento de dados de entrada.

4.2 Objetivos Específicos

Como objetivos específicos, destacamos um conjunto de etapas que compõem o processamento dos dados coletados desde a extração de características até a avaliação da solução obtida:

1. Ser capaz de trabalhar com os dados coletados de redes sociais que contam com as quatro características apresentadas na Seção 3: heterogeneidade, multimodalidade, não-rotulação e não-estruturação;

2. Aplicar engenharia de características para obter representações que possibilitem a análise e aprendizado dos dados mesmo considerando heterogeneidade, multimodalidade e não-estruturação, avaliando o desempenho de abordagens que utilizam projeção de características latentes, descrição dos dados por meio de uma única modalidade e, tratamento individual de cada modalidade;

3. Analisar técnicas de aprendizado não-supervisionado e/ou semi-supervisionado como possibilidades para a solução do problema da falta de rótulos;

4. Desenvolver métodos de validação da solução obtida com relação à relevância dos dados.

5 Metodologia

Esta seção detalha os principais passos que visamos tomar para atingir os objetivos especificados na Seção 4. A construção de um modelo de aprendizado de máquina envolve as etapas de preparação de dados, engenharia de características, aprendizado, e análise dos resultados e validação, conforme apresenta a Figura 3. Cada etapa é detalhada a seguir.

5.1 Preparação dos Dados

Considerando um conjunto de dados obtidos de redes sociais, durante a etapa de aquisição de dados no contexto do projeto *DéjàVu*, um processamento prévio deve ser realizado de forma a possibilitar a extração de características, superando problemas como redundância e falta de rótulos [36], alguns dos principais desafios desta etapa.

A redundância de dados é significativa quando estes são obtidos de redes sociais devido à replicação de notícias por parte de usuários. Utilizar amostras iguais, além de aumentar desnecessariamente o processamento, também pode ser prejudicial ao aprendizado do modelo. Metodologias mais utilizadas para eliminar duplicatas são baseadas em medidas de similaridade,

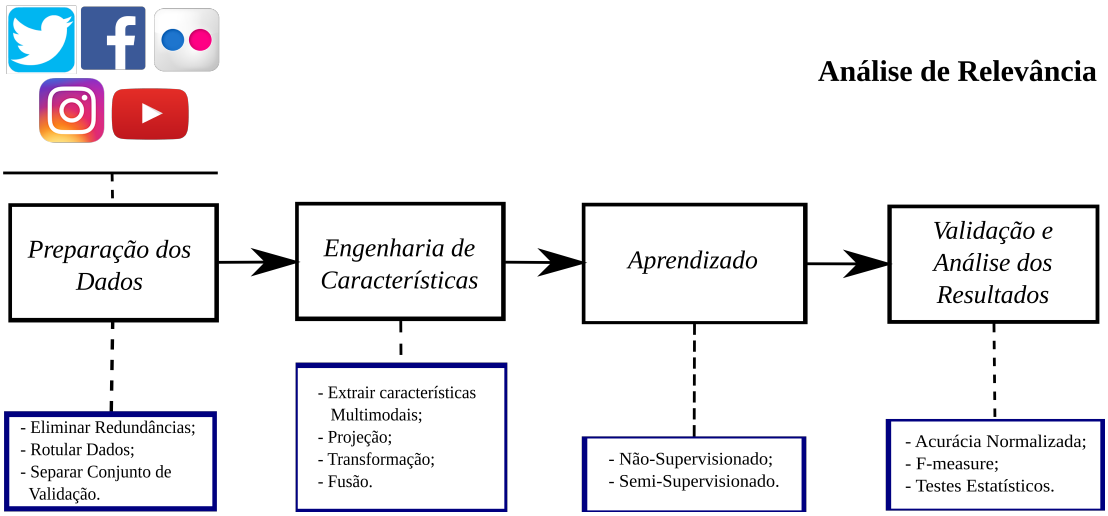

Figura 3: Etapas da Análise de Relevância. Os dados são inicialmente preparados e parte deles é rotulada para posterior validação (*Preparação dos Dados*). Na sequência, é realizada a engenharia de características que envolve a extração e modificação necessária para criação de uma representação dos dados coletados (*Engenharia de Características*), sendo esta representação utilizada pelo processo de aprendizado (*Aprendizado*) para obtenção de um modelo final que será validado no fim do processo (*Validação e Análise dos Resultados*).

como a distância euclidiana, no entanto, a comparação entre esses dados pode tornar-se cara, já que trabalharemos com conjuntos grandes de amostras.

A rotulação é essencial tanto para o direcionamento dos métodos de aprendizado quanto para a validação do modelo. Os principais desafios consistem, principalmente, na grande quantidade de dados e na necessidade de atingir grande representatividade do evento com o que está sendo rotulado, dessa maneira, essa tarefa pode ser explorada de diferentes maneiras, como por exemplo: Nguyen-Dinh *et al.* [19] fizeram uso de áudios, provenientes de repositórios criados por usuários, já rotulados, para reconhecimento de contexto, atingindo um grande número de classes e variabilidade em cada uma delas; Zhang *et al.* [34] exploraram o uso de uma pequena quantidade de dados rotulados manualmente, juntamente com dados não-rotulados, para o aprendizado semi-supervisionado de modelos em aplicações na área da saúde; e Mozarafi *et al.* [17] propuseram o uso de aprendizado ativo para combinar a capacidade humana com a velocidade dos computadores, reduzindo as questões a serem solucionadas pelo ser humano durante a rotulação e, consequentemente, aumentando o número de amostras rotuladas.

5.2 Engenharia de Características

Nesta etapa, serão extraídas as características das múltiplas modalidades para uso no modelo de aprendizado. Como descrito na Seção 3, realizaremos o tratamento da multimodalidade de três maneiras diferentes, utilizando projeção, transformação e fusão de características.

Em um primeiro momento, características individuais de cada uma das modalidades devem

ser extraídas para possibilitar posterior manipulação e tratamento multimodal. Dessa maneira, a extração será particionada em duas categorias: textual e visual.

Na categoria textual, serão explorados *embeddings* de palavras [5, 13, 30, 32]. *Embeddings* são modelos treinados para mapear palavras em vetores de características de forma a manter significados e ser possível realizar comparação e cálculo de métricas de proximidade entre palavras. Alguns trabalhos que fizeram uso de *embeddings* de palavras e que servirão de base para nosso trabalho foram apresentados por: Boom *et al.* [5] utilizando representações de sentenças pequenas para representar o significado semântico de frases fazendo uso de bases de dados do *Wikipedia* e *Twitter*; Krebs *et al.* [13] desenvolvendo uma análise de sentimentos de publicações no *Facebook* utilizando *embeddings* pré-treinados; Yang *et al.* [30] analisando as características das bases de dados utilizadas para treinamento de *embeddings* e seu impacto em uma tarefa de classificação de publicações no *Twitter* relacionadas a eleições; e Yao *et al.* [32] propondo uma abordagem de aprendizado de *embedding* dinâmico capaz de representar semântica temporal das palavras.

Na categoria visual, que agrega imagens e vídeos, serão extraídas características por meio de redes convolucionais associando, quando necessário, descritores de imagens. Redes convolucionais agregam em suas camadas escondidas filtros de processamento de imagens que possibilitam a extração de determinadas características de forma invariante a translações e escalas, por exemplo, além de incluir informações de texturas e bordas. Alguns trabalhos que utilizam CNNs como extrator de características e que podem ser usados como base foram propostos por: Garcia-Gasulla *et al.* [9] que realizaram uma análise do poder de representatividade das características geradas pelas camadas anteriores à de decisão de uma CNN; e Zhang *et al.* [35] que utilizaram a fusão de características extraídas por uma CNN com características extraídas por outros métodos de visão computacional para uma tarefa de recuperação de informação baseada em imagens.

5.3 Aprendizado

Como um grande desafio desta etapa está a necessidade de direcionar o aprendizado da relevância, não apenas para um evento em especial, mas sim de maneira generalizada, para eventos a serem definidos posteriormente. Além disso, a constante possibilidade de entrada de dados novos reforça a necessidade de um modelo dinâmico, que possibilite o uso de novas informações.

Inicialmente, serão exploradas técnicas de aprendizado não-supervisionado e como estas se comportam para diferentes eventos. A partir dos resultados obtidos, analisaremos as possíveis modificações a serem realizadas para melhorar o modelo, incluindo informação para guiar o aprendizado. É importante destacar que, possíveis informações adicionais devem ser criteriosamente escolhidas de forma a representar completamente o evento, de outra maneira, a falha na descrição pode ocasionar erros de julgamento, em especial quando novos dados são adicionados ao modelo. Outro desafio encontrado é determinar quando uma representação está suficientemente completa para a correta análise de relevância.

Como uma abordagem alternativa, que pode melhorar a performance do modelo, adicionaremos informações rotuladas ao aprendizado, explorando técnicas de aprendizado semi-supervisionado. Nessa metodologia, é importante também, garantir a qualidade das amostras rotuladas e sua representatividade, desafios da etapa de Preparação dos Dados (Seção 5.1), sendo a preocupação desta etapa determinar a melhor maneira com que os dados rotulados interagirão com os não-rotulados, de forma a proporcionar maior capacidade de generalização mantendo as taxas de aprendizado elevadas.

5.4 Validação e Análise dos Resultados

Dentre os dados rotulados, separaremos um conjunto de validação para determinação da qualidade dos resultados. Serão utilizadas métricas como *acurácia normalizada* e *F-measure*, além de testes estatísticos para avaliar os algoritmos implementados.

Quando possível, compararemos os algoritmos desenvolvidos com equivalentes na literatura. Algumas possíveis bases de dados para avaliação dos métodos incluem:

Twitter event datasets (2012-2016) : composta por uma coleção de *tweets*, relacionados com 30 diferentes bases de dados de eventos mundiais, coletados durante os anos de 2012 a 2016. A base foi apresentada e utilizada nos trabalhos de Zubiaga *et al.* [39] e Zubiaga [38];

Global Database on Events, Location and Tone (GDELT) [2]: é uma das maiores bases de dados de eventos contanto com mais de 250 milhões de eventos mundiais que aconteceram a partir de 1979, incluindo informações de data, local e atores;

Integrated Crisis Early Warning System (ICEWS) : é um conjunto de quatro bases de dados poduzidas para a *Defense Advanced Research Projects Agency* (DARPA) e o *Office of Naval Research* (ONR). Conta com a *ICEWS Coded Event Data* com eventos socio-políticos [6], a *ICEWS Event Aggregations* agrega cada evento a um intervalo de tempo [14], a *ICEWS Events of Interest Ground Truth Data Set* associa o evento de interesse à localização e data [15], e a *ICEWS Dictionaries* associa atores e agentes em cada evento descrito [7];

Phoenix Near-Real-Time Data [3]: é uma das bases de eventos mais novas mantida pela *Open Event Data Alliance (OEDA)* e construída a partir de um conjunto de notícias do *New York Times (NYT)*, *British Broadcasting Corporation's (BBC) Summary of World Broadcasts (SWB)*, e *Central Intelligence Agency's (CIA) Foreign Broadcast Information Service (FBIS)* [22];

Outras bases : algumas outras bases de dados abrangendo eventos específicos, incluem a *Armed Conflict Location & Event Data Project (ACLED)*[4] que apresenta eventos de violência e

[2] https://www.gdeltproject.org/
[3] http://eventdata.utdallas.edu/
[4] https://www.acleddata.com/

protestos na África e Ásia; e a *Global Terrorism Database (GTD)*[5] abrangendo incidentes de terrorismo pelo mundo.

Dentre as bases de dados encontradas atualmente, incluindo as aqui descritas, notamos que um desafio desta etapa será a escassez de bases multimodais, já que grande parte delas trabalha com dados textuais obtidos de notícias sobre os eventos. A partir das avaliações, escolheremos um algoritmo para compor a solução final e este será utilizado para determinação de relevância de dados de conjuntos de eventos a serem definidos. Os resultados serão avaliados manualmente, dada a inexistência de rótulos nos conjuntos de teste, de forma quantitativa e qualitativa, amostrando parte do conjunto de dados e determinando para cada dado dessa amostragem, se este foi corretamente considerado relevante ou não.

6 Plano de Trabalho

A Tabela 1 apresenta o cronograma das atividades a serem executadas no decorrer do trabalho, que são listadas a seguir:

1. Obtenção de créditos obrigatórios em disciplinas do programa de mestrado;
2. Revisão bibliográfica;
3. Escrita da proposta de mestrado;
4. Exame de Qualificação de Mestrado;
5. Participação no Programa de Estágio Docente (PED);
6. Análise e implementação da extração de características das diferentes modalidades;
7. Análise e implementação de projeções, transformações e/ou fusões de características multimodais;
8. Análise e implementação de técnicas de aprendizagem não-supervisionadas e/ou semi-supervisionadas para solução do problema;
9. Análise dos resultados;
10. Escrita/Revisão da dissertação;
11. Defesa da dissertação.

O tempo alocado e a ordem de execução de algumas atividades podem ser alterados no decorrer do desenvolvimento da pesquisa, uma vez que alguns resultados podem ser mais promissores que outros, fazendo com que algumas atividades sejam realocadas, ajustadas ou eliminadas.

[5] http://www.start.umd.edu/gtd/

	2018										2019												2020	
	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	✓	✓	✓	✓		✓	✓	✓	✓															
2		✓	✓	✓	✓	✓	✓				✓				✓				✓			✓		
3				✓	✓	✓																		
4							✓																	
5													✓	✓	✓	✓								
6							✓	✓	✓	✓														
7											✓	✓	✓	✓										
8															✓	✓	✓	✓	✓	✓	✓			
9									✓	✓			✓	✓			✓	✓		✓	✓			
10										✓				✓				✓			✓	✓	✓	
11																								✓

Tabela 1: Cronograma de atividades.

7 Resultados Preliminares

Com o intuito de iniciar as atividades de análise de eventos e compreender os desafios encontrados durante a análise de relevância, escolhemos um evento específico e uma tarefa particular, que foi executada com o uso de técnicas de aprendizado de máquina abordadas neste projeto.

O evento escolhido foi o incêndio do *Grenfell Tower*, sendo este evento registrado por pessoas nas proximidades e canais de TV, gerando um conjunto grande de imagens e vídeos que foram coletados em uma parceria entre a *Forensic Architecture*, instituto de pesquisa de Londres, e o projeto DéjàVu.

A tarefa que nos propomos a realizar foi a classificação da fachada do prédio entre norte, sul, leste e oeste, um requisito para auxiliar na compreensão do evento e construção da *Coerência-X* através de uma ordenação espacial que, em um próximo estágio, seria seguida por uma ordenação temporal. Destacamos aqui que, apesar dessa tarefa fazer parte da etapa de *Organização dos Dados e Coerência-X*, fora do escopo desta proposta, ela possibilitou a compreensão do impacto do uso de dados com diferentes níveis de relevância, assim como irrelevantes; e ressaltou ainda os desafios anteriormente apresentados (Seção 5) ao trabalhar com dados provenientes de mídias sociais e internet.

Inicialmente, a modalidade de interesse foi o conjunto de vídeos do evento, sendo: 20 videos da BBC, 1 do Metro, 5 do Periscope, 1 da Public Inquiry, 5 da Sky, 33 de mídias sociais, 11 do Storyful e 9 do YouTube. A partir dos vídeos, foram extraídos os quadros relevantes totalizando 69,468 imagens não-rotuladas. Para montar um conjunto de validação, de forma a avaliar o desempenho das abordagens, 100 imagens foram rotuladas para cada uma das quatro classes.

7.1 Aprendizado Não-Supervisionado

Supondo que suas classes correspondentes não fossem conhecidas, alguns experimentos com aprendizado não-supervisionado foram realizados com o algoritmo *K-means*. O número de fachadas é quatro, no entanto, como os ângulos de aquisição dos vídeos podem variar, testamos o algoritmo com $K = 4$, $K = 6$ e $K = 8$ clusters com sementes randomicamente escolhidas, obtendo acurácia normalizada conforme Tabela 2.

Tabela 2: Acurácia Normalizada para três diferentes números de grupos.

	4	**6**	**8**
Acc	0.38	0.55	0.56

Escolhemos $K = 6$ para a continuação dos experimentos, já que sua acurácia normalizada foi bem superior a $K = 4$ e similar a $K = 8$. Os próximos testes foram realizados escolhendo sementes iniciais em cada cluster, de acordo com a classe que desejamos compor. Foram realizados três testes com sementes iniciais diferentes. Os resultados são apresentados na Tabela 3.

Tabela 3: Acurácia Normalizada para três conjuntos diferentes de inicialização de sementes, usando dois grupos para representar as classes leste e norte.

	1	**2**	**3**
Acc	0.55	0.57	0.60

É possível perceber que a qualidade dos resultados está relacionada às sementes iniciais escolhidas, ou seja, quanto maior a relevância da imagem inicial, melhor a recuperação de imagens também relevantes. Uma maneira pensada para aumentar a robustez da clusterização foi o uso de voto majoritário realizado com os resultados de múltiplas clusterizações com conjunto de sementes diferentes escolhidas aleatoriamente. A Tabela 4 apresenta acurácia normalizada para 1, 3, 5, 7 e 9 clusterizações.

Tabela 4: Acurácia Normalizada para números diferentes de clusterizações (combinadas com voto majoritário) usando dois grupos para representar as classes leste e norte.

	1	**3**	**5**	**7**	**9**
Acc	0.57	0.62	0.54	0.58	0.58

7.2 Aprendizado Semi-Supervisionado

Para os experimentos com aprendizado semi-supervisionado, o conjunto de imagens rotuladas foi dividido ao meio em conjuntos de treinamento e validação. Além disso, foram incluídas 107 imagens não-rotuladas no conjunto de treinamento. A técnica utilizada foi a propagação de rótulos [4] que utiliza o algoritmo KNN (*K-Nerest Neighbors*) atribuindo classes por meio de proximidade com os dados rotulados. Essa técnica ainda utiliza os dados não rotulados do treinamento como uma forma de aumentar a variabilidade do modelo, ideal para quando não existem muitos dados rotulados. A Tabela 5 apresenta as métricas de acurácia normalizada e F-measure para o conjunto de validação, com diferentes valores para o parâmetro K que representa o número de vizinhos mais próximos para decisão do KNN.

Uma técnica baseada na propagação de rótulos, que apresentou melhores resultados, é o espalhamento de rótulos. A diferença entre as duas técnicas é a existência de um novo parâmetro α que indica a porcentagem de possíveis mudanças de rotulação, tornando o modelo mais robusto

Tabela 5: Acurácia Normalizada (Acc) e F-measure (F1) para o conjunto de validação utilizando números diferentes de vizinhos (K).

K	**Acc**	**F1**
1	0.63	0.62
3	0.59	0.60
5	0.53	0.50

a ruídos provocados por rotulação incorreta. A Tabela 6 apresenta as métricas de acurácia normalizada e F-measure para o conjunto de validação, com diferentes valores para o parâmetro K e α.

Tabela 6: Acurácia Normalizada (Acc) e F-measure (F1) para o conjunto de validação utilizando números diferentes de vizinhos (K) e valores para α.

K	α	**Acc**	**F1**
1	**0.1**	0.63	0.62
	0.3	0.65	0.57
3	**0.1**	0.63	0.62
	0.3	0.65	0.64
	0.5	0.64	0.64
5	**0.1**	0.64	0.62
	0.3	0.63	0.62
7	**0.1**	0.64	0.62
	0.3	0.64	0.63

7.3 Considerações Finais

Por meio destes resultados preliminares, foi possível analisar alguns aspectos do trabalho que será desenvolvido. Primeiramente, percebemos que, de fato, trabalhar somente com dados relevantes facilita o processo de análise dos eventos. A existência de dados não relacionados com o evento pode ocasionar a perda de qualidade dos modelos de aprendizado.

Por fim, notamos que indicar um conjunto de sementes iniciais ou atribuir rótulos à parte dos dados de treinamento proporciona um bom direcionamento às técnicas de aprendizado não-supervisionado e semi-supervisionado. Por outro lado, esse direcionamento deve apresentar boa e completa definição do evento, ou seja, esses dados devem refletir a relevância esperada nas imagens categorizadas como relevantes.

Referências

[1] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

[2] Flora Amato, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. Recommendation in social media networks. In *Proceedings of the 3rd International Conference on Multimedia Big Data (BigMM'2017)*, pages 213–216, Laguna Hills, California, USA, 2017. IEEE.

[3] André Arnes. *Digital Forensics*. Wiley, Norway, 1 edition, 2018.

[4] Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. *Semi-Supervised Learning*. The MIT Press, London, England, 2006.

[5] Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156, 2016.

[6] Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. ICEWS coded event data, 2018.

[7] Elizabeth Boschee, Jennifer Lautenschlager, Steve Shellman, and Andrew Shilliday. ICEWS dictionaries, 2015.

[8] Fangxiang Feng, Ruifan Li, and Xiaojie Wang. Deep correspondence restricted Boltzmann machine for cross-modal retrieval. *Neurocomputing*, 154:50–60, 2015.

[9] Dario Garcia-Gasulla, Ferran Parés, Armand Vilalta, Jonatan Moreno, Eduard Ayguadé, Jesús Labarta, Ulises Cortés, and Toyotaro Suzumura. On the behavior of convolutional nets for feature extraction. *Journal of Artificial Intelligence Research*, 61:563–592, 2018.

[10] Gregor Große-Bölting, Chifumi Nishioka, and Ansgar Scherp. Generic process for extracting user profiles from social media using hierarchical knowledge bases. In *Proceedings of the 9th International Conference on Semantic Computing (ICSC'2015)*, pages 197–200, Anaheim, California, USA, 2015. IEEE.

[11] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3-4):248–260, 2009.

[12] Bongjun Kim and Bryan Pardo. A human-in-the-loop system for sound event detection and annotation. *ACM Transactions on Interactive Intelligent Systems*, 8(2):1–23, 2018.

[13] Florian Krebs, Bruno Lubascher, Tobias Moers, Pieter Schaap, and Gerasimos Spanakis. Social emotion mining techniques for Facebook posts reaction prediction. In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART'2018)*, volume 2, pages 211–220, Funchal, Madeira, Portugal, 2018.

[14] Jennifer Lautenschlager, Steve Shellman, and Michael Ward. ICEWS event aggregations, 2015.

[15] Ian Lustick, Sean O'Brien, Steve Shellman, Timothy Siedlecki, and Michael Ward. ICEWS events of interest ground truth data set, 2015.

[16] Derlin Morocho, Aythami Morales, Julian Fierrez, and Javier Ortega-Garcia. Humans in the loop: Study of semi-automatic signature recognition based on attributes. In *Proceedings of the 51st International Carnahan Conference on Security Technology (ICCST'2017)*, pages 1–5, Madrid, Spain, 2017.

[17] Barzan Mozafari, Purna Sarkar, Michael Franklin, Michael Jordan, and Samuel Madden. Scaling up crowd-sourcing to very large datasets: A case for active learning. In *Proceedings of the 41st International Conference on Very Large Data Bases (VLDB'2015)*, volume 8, pages 125–136, Kohala Coast, Hawai, 2015.

[18] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML'2011)*, pages 689–696, Bellevue, Washington, USA, 2011. Omnipress.

[19] Long-Van Nguyen-Dinh, Mirco Rossi, Ulf Blanke, and Gerhard Tröster. Combining crowd-generated media and personal data: semi-supervised learning for context recognition. In *Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia (PDM'2013)*, pages 35–38, Barcelona, Spain, 2013.

[20] Klimis Ntalianis and Nicolas Tsapatsoulis. Multiresolution organization of social media users' profiles: Fast detection and efficient transmission of characteristic profiles. In *Proceedings of the 10th IEEE International Conference on Cyber, Physical and Social Computing (CPSCom'2017)*, pages 37–43, Exeter, UK, 2017. IEEE.

[21] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, and Jie Shao. Multi-modal event topic model for social event analysis. *IEEE Transactions on Multimedia*, 18(2):233–246, 2016.

[22] Althaus Scott, Joseph Bajjalieh, John F. Carter, Buddy Peyton, and Dan A. Shalmon. Cline center historical Phoenix event data. v.1.0.0, 2017.

[23] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. The spread of fake news by social bots. *arXiv:1707.07592*, 2017.

[24] Xiaoxiao Shi and Philip S. Yu. Heterogeneous embedding via aggregating multiple sources. *Annals of Data Science*, 1(1):73–93, 2014.

[25] Vivek K. Singh, Mingyan Gao, and Ramesh Jain. Situation recognition: an evolving problem for heterogeneous dynamic big multimedia data. In *Proceedings of the 20th ACM International Conference on Multimedia (ACMMM'2012)*, pages 1209–1218, Nara, Japan, 2012. ACM New York.

[26] Hazem Souid, Chiraz Trabelsi, Gabriella Pasi, and Sadok Ben Yahia. Hypergraph fuzzy minimals transversals mining: A new approach for social media recommendation. In *Proceedings of the 26th International Conference on Fuzzy Systems (FUZZ-IEEE'2017)*, pages 1–7, Naples, Italy, 2017. IEEE.

[27] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *Proceedings of the 29th Representation Learning Workshop (ICML'2012)*, pages 1–8, Edinburgh, Scotland, UK, 2012.

[28] Wei Wang, Xiaoyan Yang, Beng Chin Ooi, Dongxiang Zhang, and Yueting Zhuang. Effective deep learning-based multi-modal retrieval. *The International Journal on Very Large Data Bases*, 25(1):79–101, 2015.

[29] Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. StarSpace: Embed all the things! *arXiv:1709.03856*, 2017.

[30] Xiao Yang, Craig Macdonald, and Iadh Ounis. Using word embeddings in Twitter election classification. *Information Retrieval*, 21:1–25, 2017.

[31] Xiaoshan Yang, Tianzhu Zhang, Changsheng Xu, and M. Shamim Hossain. Automatic visual concept learning for social event understanding. *IEEE Transactions on Multimedia*, 17(3):346–358, 2015.

[32] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM'2018)*, pages 673–681, Los Angeles, California, USA, 2018.

[33] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2016)*, pages 5534–5542, Las Vegas, NV, USA, 2016. IEEE.

[34] Gang Zhang, Shan-Xing Ou, Yong-Hui Huang, and Chun-Ru Wang. Semi-supervised learning methods for large scale healthcare data analysis. *International Journal of Computers in Healthcare*, 2(2):98–110, 2015.

[35] Guixuan Zhang, Shuwu Zhang, Zhi Zeng, Hu Guan, and Fangxin Wang. Region based image retrieval with query-adaptative feature fusion. In *Proceedings of the 24th IEEE International Conference on Image Processing (ICIP'2017)*, pages 3675–3679, Beijing, China, 2017.

[36] Lina Zhou, Shimei Pan, Jianwu Wang, and Athanasios V. Vasilakos. Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 237:350–361, 2017.

[37] Shixiang Zhu and Yao Xie. Crime event embedding with unsupervised feature selection. *arXiv:1806.06095*, 2018.

[38] Arkaitz Zubiaga. A longitudinal assessment of the persistence of Twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8):974–984, 2018.

[39] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29, 2016.