Serviço de Detecção e Enriquecimento de Eventos Rodoviários Baseado em Fusão de Dados Heterogêneos para VANETs

Paulo H. L. Rettore[1], Ígor Araújo[2], Guilherme Maia[1], Leandro A. Villas[3], Antonio A. F. Loureiro[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brasil

[2]Departamento de Engenharia
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brasil

[3]Instituto de Computação, Universidade de Campinas (UNICAMP)
Campinas, SP – Brasil

{rettore,jgmm,loureiro}@dcc.ufmg.br,

igoral@ufmg.br,leandro@ic.unicamp.br

Abstract. *In this work, we present the Twitter Incident (T-Incident), a low-cost learning-based road incident detection and enrichment approach built using heterogeneous data fusion. We design a spatiotemporal grouping model to fuse incident data, not-incident data, and Location-Based Social Media (LBSM) data. Moreover, we filter the LBSM data source using refined methods of Natural Language Processing (NLP) to extract patterns on social media data that may describe the incident event and its surrounding. After that, we used a learning-based model to identify these patterns and detect the event types. The methodology results show the best set of parameters that can be used feed our approach, achieving scores above 90% in F1 sore, Recall and Precision metrics. Allowing the incident detection and its description as T-Incident' services.*

Resumo. *Este trabalho apresenta o T-Incident, uma arquitetura robusta de baixo custo para detecção e enriquecimento de eventos rodoviários baseado na fusão de dados heterogêneos. Foi desenvolvido um modelo espaço-temporal para fusão de dados de incidente, não-incidente e mídia social. Além disso, filtrou-se essa última fonte de dados usando métodos de processamento de linguagem natural para detecção de padrões capazes de descrever o evento e sua vizinhança. Também foi desenvolvido um modelo baseado em aprendizagem para identificar esses padrões e detectar os tipos de eventos. Os resultados da metodologia mostraram os melhores parâmetros para a abordagem T-Incident, fornecendo um serviço apurado de detecção e descrição de incidentes acima de 90% para as métricas F1 score, Recall e Precisão.*

1. Introdução

O planejamento e gestão de sistemas de transporte têm se tornado crucial para promover o crescimento das cidades. Desse modo, governos, pesquisadores e indústrias [Bazzan and Klügl 2013] têm investido esforços com o intuito de compreender a mobilidade

urbana e de desenvolver soluções para a redução de congestionamentos e incidentes rodoviários. Nesse sentido, emerge o conceito de Sistemas de Transporte Inteligentes – Intelligent Transportation Systems (ITSs). Que por meio de tecnologias de informação e comunicação (Redes Veiculares *Ad-hoc* – Vehicular Ad-hoc Networks (VANETs)), visam melhorar o processo de tomada de decisão e a concepção de aplicações e serviços para os sistemas de trasporte. Entretanto, ITSs dependem de uma vasta cobertura de dados e comunicação, como dados de semáforos, fluxo de tráfego, eventos rodoviários, medidores de velocidade, dados de sensores veiculares, sensores meteorológicos, mídias entre outros. Além disso, o acesso a maioria desses dados limita a análise de tráfego em tempo real, uma vez que, sua disponibilidade muitas vezes é desatualizada, ou até mesmo apenas um conjunto limitado de empresas detêm o controle sob esses dados. Existem muitas razões para que isso ocorra, dentre elas destacamos o valor comercial dos dados e a infraestrutura precária que afetam a qualidade do sensoriamento e comunicação desses dados.

Logo, a informação entregue para os usuários, como eventos rodoviários, contém pouca descrição ou é desatualizada. Portanto, diminui a eficiência no gerenciamento de rotas de veículos, controle de fluxo e a disseminação de descrições detalhadas e úteis de um dado evento. Solucionar essas questões, melhorando a eficiência e qualidade da mobilidade, envolve conhecimento multidisciplinar. Para prover mais consistência, acurácia e informações úteis, a integração de múltiplas fontes de dados é uma tarefa essencial. Esse processo é denominado Fusão de Dados e constitui uma atividade desafiadora, dado a heterogeneidade das fontes de dados, o não sincronismo, e a existência de ruídos e erros nos dados. Além disso, o armazenamento dos dados e seu aspecto espaço-temporal contribui para a complexidade no processo de fusão desses dados heterogêneos [Khaleghi et al. 2013b].

Nesse sentido, a fim de melhorar as formas atuais de detecção e descrição de incidentes, foi desenvolvido o Twitter Incident (T-Incident), uma arquitetura robusta de baixo custo para detecção e enriquecimento de incidentes rodoviários, construído a partir da fusão de dados heterogêneos de um ITS. T-Incident traz para os sistemas de navegação (serviços de rota, detecção e descrição de eventos), planejadores de trânsito e o público em geral, uma plataforma consistente, precisa e com informações enriquecidas sobre os incidentes rodoviários. Essa arquitetura vai em direção do aprimoramento dos sistemas de navegação, introduzindo a fusão de dados como auxílio para que os usuários possam escolher suas rotas e gestores possam planejar e controlar o fluxo de veículos nas vias urbanas.

Inicialmente, foi criado um modelo de agrupamento espaço-temporal para fundir duas fontes de dados, Here WeGO[1] e Bing Maps[2], fornecendo uma *Camada de Incidentes*. Usando este mesmo modelo e inserindo dados relacionados a pontos turísticos (adquiridos da plataforma TripAdvisor[3]), dados de Mídia Social Baseada em Localização – Location-Based Social Media (LBSM) – (adquiridos do *Twitter*[4]) e a *Camada de Incidentes*, agrupou-se esses dados e utilizou-se métodos refinados de Processamento de Lingua-

[1] `https://wego.here.com`
[2] `https://bing.com/maps`
[3] `https://tripadvisor.com/`
[4] `https://developer.twitter.com/en/docs`

gem Natural – Natural Language Processing (NLP), com o objetivo de detectar padrões que descrevam o evento incidente e seu entorno. Baseado em modelos de aprendizagem de máquina foi mapeado esses padrões e detectado os tipos de eventos. Os resultados da metodologia mostram os melhores parametros para a abordagem T-Incident, fornecendo um serviço apurado de detecção de incidentes acima de 90% com as métricas *F1 score*, *Recall* e *Precisão*.

O artigo está organizado da seguinte forma. Na Seção 2, apresenta-se as abordagens de detecção de eventos que fazem uso de LBSM. A Seção 3 descreve o processo de aquisição de dados. Já na Seção 4, apresenta-se a fusão de dados com o intuito de enriquecer a cobertura de eventos incidentes. Explica-se a arquitetura T-Incident na seção 5, e a avaliação na Seção 6. Finalmente, na Seção 7, destacam-se as observações finais e trabalhos futuros.

2. Trabalhos Relacionados

O crescimento da internet e das LBSMs permitem uma vasta gama de investigações devido a quantidade de dados disponíveis diariamente. Nesta perspectiva, vários estudos foram realizados para analisar as condições de tráfego usando Mídia Social [Xu et al. 2018]. Santos et al. [Santos et al. 2018] argumenta que o uso de LBSM pode oferecer uma nova camada para melhorar a compreensão de tráfego e do trânsito. Eles também apresentaram o Twitter MAPS (T-MAPS), um modelo espaço-temporal de baixo custo para descrição das condições de tráfego usando *tweets*.

Yazici et al. [Yazici et al. 2017] realizou uma análise do *Twitter* para detectar eventos com base em contas organizacionais (ou especializadas) e regulares. Esse trabalho mostra que *tweets* coletados de contas regulares são mais propensos para ser irrelevantes, embora possam capturar eventos que acabaram de acontecer. Por outro lado, *tweets* de contas especialistas são mais valiosos e estruturados, sendo melhores para identificar eventos de incidente. Eles mostram que a combinação de ambas as fontes leva a um melhor resultado ao lidar com a detecção de eventos. Nesse sentido, Zhang et al. [Zhang et al. 2018] complementam o cenário de detecção de incidentes usando dados de mídia social. Eles mostraram que dados de mídia social podem ser úteis como uma maneira alternativa de melhorar métodos tradicionais na detecção de tráfego em tempo real.

Nguyen et al. [Nguyen et al. 2016] desenvolveram o *TrafficWatch*, um sistema de tempo real baseado em *Twitter* que visa enriquecer informações relacionadas ao tráfego para análise e visualização de incidentes na Austrália. Eles também desenvolveram um estudo de caso para detectar incidentes na estrada antes do Centro de Gerenciamento de Transporte – Transport Management Centre (TMC), e detectaram aqueles que não são relatados por ele. Pereira et al. [Pereira et al. 2013] fizeram uso de uma mídia confiável disponibilizada por centros de gerenciamento de tráfego, apresentando a modelagem de tópico (*topic modeling*), uma análise de texto para melhorar a precisão na medição de tempos de duração de um incidente. Eles provaram que o uso dessa análise melhora a previsão de incidente em 28% em vez da sua não utilização.

Neste sentido, este trabalho propõe uma arquitetura robusta para lidar com contas regulares e especializadas do *Twitter*, como em [Yazici et al. 2017], fornecendo serviços de detecção e enriquecimento de incidentes rodoviários auxiliando os usuários em suas escolhas de rotas. Percebe-se que Nguyen et al. [Nguyen et al. 2016] não descrevem

como os *tweets* e os relatórios de TMC foram correlacionados espaço-temporalmente. Por outro lado, este trabalho descreve detalhadamente a metodologia aplicada, como o modelo de agrupamento espaço-temporal, o processo de extração de características, além de utilizar fontes de dados acessíveis para classificar os eventos. Ao contrário de [Pereira et al. 2013], aqui utiliza-se uma fonte de dados não confiável, o *Twitter*, mais um conjunto de técnicas e outras fontes de dados que se mostraram muito promissoras.

3. Aquisição de Dados

A falta de informação no ambiente urbano é um dos desafios mais importante em ITS. Portanto, investigações desse contexto são restritos a estudos teóricos ou baseados na vasta quantidade de dados privados. Felizmente, o aumento do uso de plataformas online, como as LBSM, fazem com que seja possível as pessoas compartilharem seus dados, vida cotidiana e opiniões em uma variedade de campos, inclusive situações de trânsito. Partindo dessa perspectiva, foram coletados dados do *Twitter*, plataforma online na qual pessoas frequentemente compartilham informações. Foi utilizado sua API como método de coleta de informações do usuário, respeitando os termos de restrição da empresa.

Em um trabalho anterior, T-MAPS, um modelo espaço-temporal de baixo custo para melhorar a descrição das condições de tráfego através de *tweets* [Santos et al. 2018] foi usado como um guia para a abordagem aqui utilizada devido à comprovação da viabilidade de se estudar e explorar a relação da LBSM com o cenário de tráfego. O T-Incident é uma abordagem para identificar com precisão e eficiência os eventos de tráfego (incidentes e não-incidentes) e também enriquecer suas descrições. O processo de aquisição de dados visa combinar diferentes fontes de dados, como Here WeGo, Bing Maps, Tripadvisor e *Twitter* nas dimensões temporais e espaciais para alcançar esses objetivos.

O conjunto de dados utilizados neste trabalho é constituído de 158.413 *tweets* adquiridos via *streaming* aberto entre os dias 14-09-2018 e 06-11-2018. Um conjunto de palavras relacionadas a incidentes como *congestion, accident, construction, planned event, road hazard, disabled vehicle, traffic, jam, car, weather*, foram utilizadas para delimitar o escopo. Todos os *tweets* tinham geolocalização e, em grande parte, estavam localizados em Manhattan - Nova York. Neste trabalho, estamos interessados em *tweets* de usuários regulares e especialistas (contas controladas por empresas). Ademais, descartou-se os *tweets* postados como *retweet*, ou seja, o interesse aqui está em coletar impressões do usuário e não a propagação de informações.

Os dados de LBSM apresentam aspectos importantes que precisam ser considerados antes de seu processamento [Rettore et al. 2016, Santos et al. 2018, Khaleghi et al. 2013a]. Dentre os problemas dos dados descritos pelos autores, destacam-se alguns que tiveram que ser tratados neste trabalho, como: *Imprecisão de dados*, *Viés do usuário*, *Lacunas Espaço-temporais* e *Inconsistências*.

Para coletar eventos de incidentes, tanto quanto possível, desenvolveu-se uma abordagem para adquirir e tratar os dados de duas fontes de dados, Here WeGo e Bing Maps, fazendo uso de suas respectivas APIs. A área de interesse e o intervalo de tempo foram os mesmo do processo de aquisição de dados do *Twitter*. Os incidentes fornecidos pelas plataformas foram coletados por hora, resultando em 9.784 incidentes distintos adquiridos do Here WeGo e 1.924 incidentes distintos adquiridos do Bing Maps. Para utilizar esses incidentes aplicou-se a fusão dessas duas fontes de dados aumentando a

Tabela 1: Aquisição de dados.

Fonte	Objetvo	Amostra	Tempo	Espaço
Twitter	Ponto de Vista	158413	14-09-18 06-11-18	Manhattan New York
Here WeGo	Incidente	9784		
Bing Maps	Incidente	1924		
Trip Advisor	Não-incidente	50		

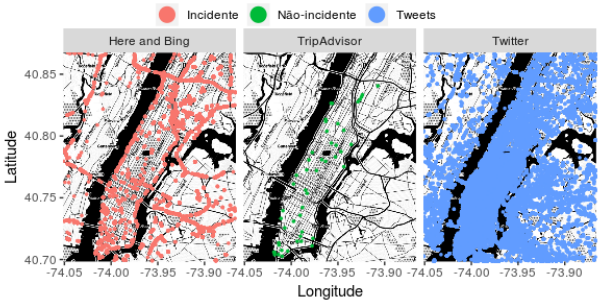


Figura 2: Cobertura espacial das fontes de dados.

cobertura e enriquecendo os incidentes semelhantes (a Seção 4 detalha esse processo).

Visando detectar incidentes, também é necessário compreender o que é um não incidente. Baseado nisso, escolheu-se locais sem evidência de incidentes, coletando dados de fontes que lidam com locais turísticos, como o Tripadvisor, um site de viagens que indica shows, lugares, hotéis, comentários sobre restaurantes, e outros conteúdos relacionados a viagem. Portanto, um conjunto dos lugares mais populares avaliados por turistas foi escolhido, como museus, parques, pubs, teatros, dentre outros. A Tabela 1 resume a aquisição de dados conduzida neste trabalho, e a figura 2 mostra a cobertura de dados espacial, por cada fonte de dado, usada para desenvolver a abordagem do T-Incident.

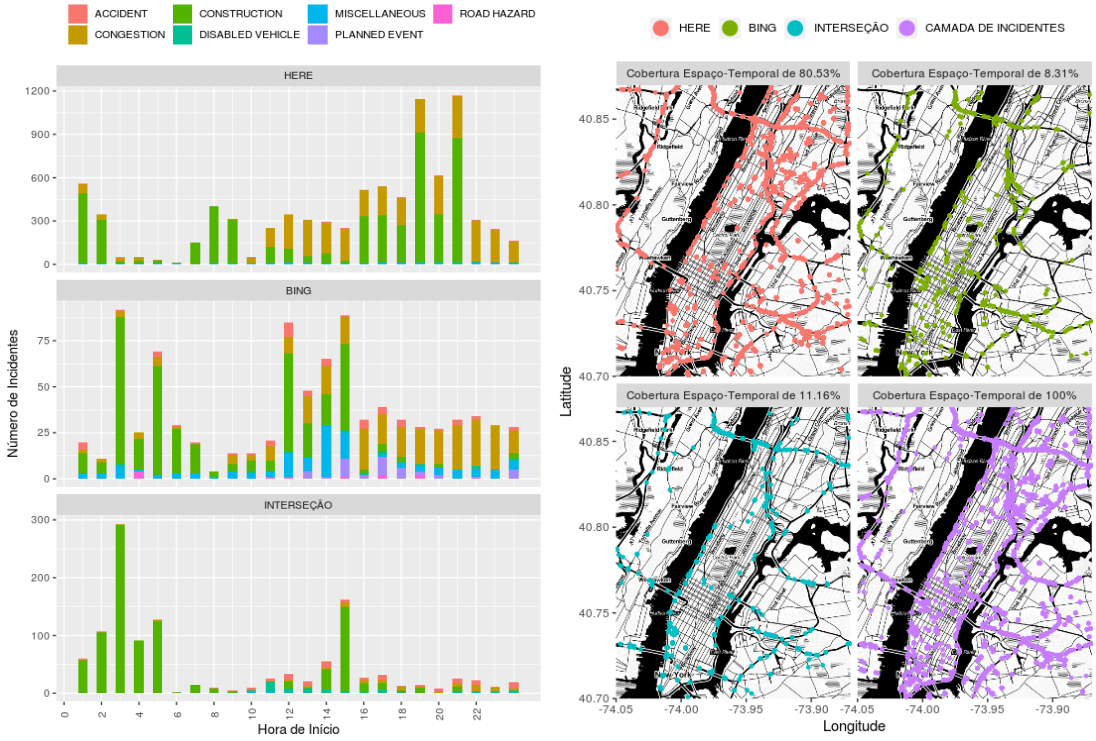
4. Fusão de Dados de Incidentes

Nesta Seção, é apresentado um método para melhorar a cobertura de dados de incidentes, fundindo diferentes fontes de dados. Esse processo existe para que mais agrupamentos possam ser feitos com os *tweets*, melhorando a abordagem proposta. Nesse sentido, coletou-se dados das plataformas *Here WeGo* e *Bing Maps*, pré-processamos para padronizar a linguagem que ambas as fontes de dados usam.

Em seguida, foi criado um modelo espaço-temporal (veja Seção 5.1 para mais detalhes, incluindo o Algoritmo 1), com o objetivo de identificar incidentes similares. Desse modo, o intervalo temporal e a localização espacial devem ser próximas. Ou seja, eventos similares devem começar no mesmo dia e hora e estar localizado em um raio máximo de 10 metros entre si. A Figura 3a mostra a frequência de cada tipo de incidente por determinada fonte de dados. Além disso, é possível ver o mesmo evento reportado por ambas as fontes no gráfico de *Interseção*.

Nesse sentido, avaliou-se também a similaridade de tipos de incidentes na interseção entre as duas fontes de dados. Descobrindo-se que a similaridade de incidentes na interseção entre *Here* e *Bing* alcançou 99.83%, evidenciando a qualidade dos dados adquiridos. Por fim, foi fornecido uma nova *Camada de Incidentes*, que resulta em uma maior cobertura e descrição das informações sobre o incidente, uma vez que, cada fonte tem sua forma de reportar e, portanto, a combinação delas enriquece o contexto atual.

A Figura 3b mostra a cobertura espacial de cada fonte de dados e a interseção entre elas durante o processo de aquisição de dados. Ou seja, considerando *Here WeGo* como H e *Bing Maps* como B, temos $(H \cup B)$ = 100% dos dados, H = 80,53% , B = 8,31% e $(H \cap B)$ = 11,16%. A seguir, exibe-se a nova *Camada de Incidentes* que cobre 100%



(a) Hora de início dos incidentes. (b) Cobertura espacial dos incidentes.

Figura 3: Cobertura espaço-temporal dos incidentes por fonte de dados.

de todos os dados coletados, sendo que mais de 11% de eventos semelhantes puderam ser enriquecidos com informações mais detalhadas.

5. Arquitetura do T-Incident

A abordagem de detecção de incidente baseada em fusão de dados heterogêneos foi conduzida considerando a premissa que as LBSM podem fornecer informações valiosas a respeito do tráfego e incidentes. Essa premissa foi parcialmente respondida em um trabalho anterior [Santos et al. 2018].

Desse modo, projetou-se um modelo de agrupamento espaço-temporal que tem como objetivo combinar diferentes fontes de dados em dimensões temporais e espaciais. Em seguida, conduziu-se um processo de extração de características, com o objetivo de identificar os pontos de vista dos usuários em torno do evento rodoviário. Então, foi desenvolvido um modelo baseado em aprendizado de máquina para identificação de incidentes considerando os relatos de usuários de LBSM. Finalmente, avaliou-se a abordagem proposta usando diferentes modelos de agrupamento espaciais. Cada estágio da abordagem T-Incident é apresentado na Figura 4.

5.1. Modelo de Agrupamento Espaço-temporal

O modelo de agrupamento leva em consideração a heterogeneidade dos dados e suas variações na cobertura espacial e temporal. Portanto, foi proposto uma abordagem que sobrepõe as camadas de dados de incidentes/não-incidentes com a camada de *tweets* considerando ambas as dimensões. Para isso, cada tipo de incidente foi agrupado como um

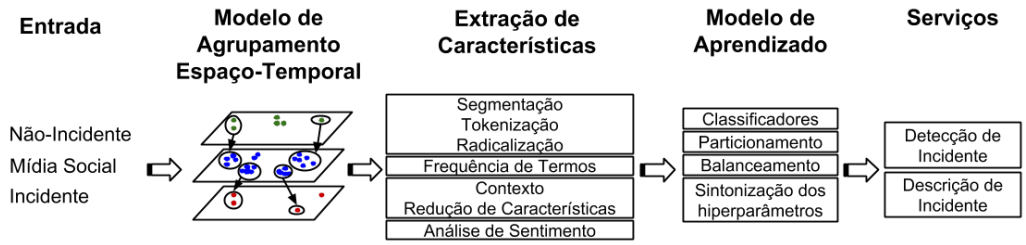


Figura 4: Arquitetura do T-Incident.

único evento – *Incidente*. Cada incidente tem uma duração e localização de início e fim. Este modelo usa somente a localização inicial dos eventos, seja ele *Incidente* ou *Não-incidente*. Foi definido também o intervalo de tempo dos *Não-incidentes* como sendo o mesmo do início e fim da coleta dados do *Twitter*.

Baseado no conjunto de dados de incidentes e não-incidentes, foi aplicado um filtro temporal para encontrar a intersecção entre eventos e *tweets*. Uma vez que as intersecções foram encontradas, o filtro espacial baseado em raios ajustáveis em torno de cada evento foi criado (0,01 km, 0,05 km e entre 0,1 e 0,5 km com intervalo de 100 m) visando, assim, identificar o melhor agrupamento, uma vez que é preciso lidar com o viés do usuário e, sobretudo, um vasto conjunto de dados não relacionados. Essa metodologia permite agrupar diferentes números de *tweets* ao redor do evento (ver a Tabela 1), e como consequência, a informação acerca do evento pode ser mais específica ou mais generalista para o contexto.

Tabela 1: Número de tweets por cada modelo de agrupamento espaço-temporal.

Evento	**Raios (km)**						
	0.01	0.05	0.1	0.2	0.3	0.4	0.5
Incidente	121	959	3098	9467	30085	63853	68877
Não-incidente	260	3161	6522	13060	20699	30492	35786

Mesmo que o agrupamento espaço-temporal possa ser conduzido de diferentes formas, como baseado em segmentos de rua, vizinhanças e, também, dividindo uma área geográfica, por exemplo. Foi definido para este trabalho o modelo de agrupamento por raio ajustável ao redor do incidente. *Tweets* não agrupados, ou seja, que ficaram fora do alcance dos raios, foram denominados como desconhecidos e removidos das etapas posteriores. Nesse processo, percebeu-se um *trade-off* na escolha dos raios, sendo os muito pequenos implicando em poucos dados agrupados, porém com informações relevantes acerca do evento. Já os raios maiores agrupam maiores volumes de dados, entretanto com menos informações que descrevem o evento. Esta situação tornou-se um desafio quando a quantidade de dados adquiridos é reduzida.

O Algoritmo 1 descreve o modelo de agrupamento espaço-temporal aplicado neste trabalho. As entradas para o modelo são *tweets*, incidentes (e não-incidentes) e os raios. O resultado esperado são os *tweets* com seus respectivos incidentes agregados (id e tipo do incidente). Foi desenvolvido também um processo otimizado que fragmenta a área geográfica, latitudinalmente, em x seções, visando a redução da computação em grandes áreas com grandes quantidades de dados. Nesse sentido, para cada *tweet* e incidente,

Algorithm 1: Modelo de agrupamento espaço-temporal

```
   Entrada: tweets,incidentes,raios
   Result: tweets agrupados por evento, Id do incidente, and Tipo de incidente
   /* Data-set divido em x pedaços, reduzindo a computação                    */
1  initialization;
2  for cada tweets do
3      currentIncidentId ← 0;
4      currentIncidentTmp ← None;
5      currentDistance ← ∞;                                  /* maior valor que o raio */
6      for each incidents do
7          if equal(tweets.sec,incidents.sec) or diff(tweets.sec,incidents.sec) is (+ 1 or - 1) then
               /* Tweets contidos no intervalo do incidente                   */
8              if TemporalFilter(incidents.starttime, incidents.endtime, tweets.timestamp) then
                   /* Distância do raio                                       */
9                  distance ← SpatialFilter(tweets.coord, incidents.coord, currentDistance, radius);
                   /* Grava a menor distaância                                */
10                 if distance < currentDistance then
11                     currentIncidentId ← incidents.Id;
12                     currentIncidentTmp ← incidents.Type;
13                     currentDistance ← distance;
14                 end
15             end
16         end
17     end
       /* Atribui tipo de incidente (INCIDENTE,NÃO-INCIDENTE) para cada tweet  */
18 end
```

é avaliado se estão na mesma seção ou perto (uma seção acima ou abaixo) (linha 7). Satisfeita essa condição, o *tweet* também deve estar entre o início e fim do incidente (linha 8). Logo, a distância entre o *tweet* e o incidente é calculada, buscando encontrar a menor distância para que a atribuição seja feita (linha 9-14).

5.2. Extração de Características

Considerando que as informações de interesse estão ao redor do local observado, é importante enfatizar que o raio é uma abordagem poderosa e intuitiva, como provado na Seção 6. Contudo, os dados coletados da LBSM contém problemas que podem levar a outros desafios como, por exemplo, a imprecisão dos dados e o viés dos usuários. Nesse sentido, o papel da extração de características é filtrar os *tweets* e prover um conjunto de palavras que melhor descrevem as redondezas do evento.

Primeiramente foi conduzido para cada modelo e classe de evento um conjunto de métodos de NLP, como transformação para palavras minúsculas, remoção de acentos, extração de sinal, filtro de *stop words*, links e caracteres especiais. Logo depois, foram reduzidas as formas inflexionais e derivacionais de cada palavra para uma forma básica comum. Assim, analisou-se a Frequência dos Termos – Term Frequency (TF) – para cada evento, extraindo uma matriz com as palavras que foram mais frequentes nessa área. Além disso, filtrou-se essa matriz baseada na matriz de dispersão, ou seja, removeu-se termos que fossem mais esparsos que 0.98%.

Também foi introduzido o passo denominado *Contexto*, cujo objetivo é a intervenção de um especialista no instante em que o conjunto de palavras processadas lhe é entregue, automaticamente pela abordagem, para identificação de relevância para o contexto, sendo mantidas apenas as palavras relacionadas ao contexto observado. Esta etapa é necessário, pois as LBSMs contém ruídos que precisam ser removidos. O final desse processo resultou em um subconjunto das palavras mais relevantes de cada categoria de incidente e raio. A Figura 5 mostra o exemplo de agrupamento por raios entre 0,01 e 0,5 km. Indicando o quão específica ou geral pode ser a informação em torno do evento em relação ao raio. As Figuras 5a e 5b apresentam mais palavras, com frequências

diferentes, o que reduz a intersecção entre os conjuntos incidentes e não-incidentes. Contudo, ao aumentar o raio é possível perceber menos palavras, com frequências próximas, Figuras 5c e 5d, enfatizando a semelhança entre ambas as classes.

(a) Tweets nas áreas de incidentes.

(b) Tweets fora das áreas de incidentes.

(c) Tweets nas áreas de incidentes.

(d) Tweets fora das áreas de incidentes.

Figura 5: Modelo espaço-temporal com raio de 0,01 km ((a) e (b)) e 0,5 km ((c) e (d)).

5.2.1. Redução de Características

O número de características (palavras) obtidas no último estágio pode ser grande o suficiente para introduzir barreiras computacionais. Desse modo, foi realizado um método para redução do número de características baseado na sua menor importância ou menor frequência. Inicialmente, desenvolveu-se duas abordagens para atingir esta meta. A primeira foi o uso da Análise de Componentes Principais – *Principal Component Analysis (PCA)* – para extração de conjuntos de características relevantes. Este processo identifica a maior variabilidade das características de um conjunto de dados e expressa em Componentes Principais (CPs). Esses CPs representam as direções pelas quais a variação nos dados é máxima. A segunda abordagem é baseada na classificação de palavras mais frequentes.

Ambos os métodos geram resultados para o especialista decidir a melhor opção. Percebeu-se que, quando os *tweets* são coletado sem palavras-chave (*track*), o PCA conseguiu um desempenho melhor que o utilizado com as palavras mais frequentes. Por outro lado, o PCA não se comportou bem em *tweets* com palavras específicas, como exposto na Seção 3. Como resultado, foi reduzida as características dos dados para cada modelo de agrupamento e classe de evento, extraindo somente o conjunto de palavras mais representativas. A Tabela 2 mostra um exemplo de características obtidas após a classificação das palavras mais frequentes no modelo de agrupamento de 0,01 km.

Tabela 2: Características relevantes baseado na raio de 0.01 km.

Eventos	Característica mais frequentes					
Incidentes	traffic	side	exit	accid	incid	clear
	avenu	contruct	updat	georgewashingtonbridg	level	street
	jersey	event	major	franklindrooseveltdr		
Não-incidentes	town	night	year	apollotheat	show	talk
	rockwood	hall	music	halloween	detail	stage
	week	play	live			

5.2.2. Análise de Sentimentos

A análise de sentimentos tem o objetivo de aumentar a descrição dos eventos. Esse processo foi conduzido para cada *tweet* em cada modelo de agrupamento e classe de evento, permitindo extrair os sentimentos dos usuários sobre o evento que eles presenciaram. Esse análise permite criar novos descritores para os eventos, ou seja, a análise de sentimento permite adicionar um conjunto de características aos eventos que posteriormente alimentarão o modelo de aprendizagem, melhorando o processo de classificação. Para extração dos sentimentos, foi usado um dicionário de palavras e seus sentimentos associados [Jockers 2017]. Os sentimentos dependem do número de palavras/sentimentos para calcular a pontuação e assim ser associado um sentimento (positivo ou negativo) ao *tweet*. Como resultado, para cada *tweet* foi extraído um conjunto de palavras de sentimentos e suas frequências. Em seguida, para esse mesmo *tweet*, agrupou-se também o conjunto de palavras obtido no estágio anterior.

5.3. Modelo baseado em aprendizagem

O modelo baseado em aprendizado é alimentado com o conjunto de características previamente rotuladas para cada evento e agrupamento. Inicialmente, foi escolhido os algoritmos de classificação (*kernels*) mais explorados pela literatura para resolução de problemas nesse contexto [Xu et al. 2018], ou seja, Support Vector Machine (SVM), k-Nearest Neighbors (KNN) e Random Forest Classifier (RF).

Em seguida, particionou-se os dados em dois conjuntos, cada modelo de agrupamento, seguindo a convenção da maioria das abordagens de aprendizagem de máquina. Sendo o conjunto de treinamento, correspondendo a 70% do conjunto de dados, e 30% correspondendo ao conjunto de teste. Para validar o treinamento, foi aplicada a esta base a validação-cruzada, considerando 10 partições divididas também em 70% e 30% de treinamento e validação, respectivamente. O objetivo desse passo é avaliar as curvas de treinamento e as curvas de testes, evitando possíveis *over-fitting* e *under-fitting*.

Percebeu-se que o conjunto de dados apresenta desbalanceamento entre as classes, uma vez que o número de *tweets* em torno das áreas de incidentes e não-incidentes pode variar. Na existência de técnicas para lidar com esse problema, explorou-se as técnica de re-amostragem que visam equilibrar as classes aumentando a frequência da classe minoritária ou diminuindo a frequência da classe majoritária. Desse modo, foi usada uma subamostragem aleatória, com o objetivo de equilibrar a distribuição de classes escolhendo aleatoriamente e eliminando os exemplos da classes majoritária.

O passo seguinte consistiu em encontrar os melhores hiper-parâmetros para os

algoritmos executados e uma abordagem exploratória foi adotada para lidar com esse processo. Usou-se, então, o método *GridSearchCV*, uma classe do Scikit-Learn API [Pedregosa et al. 2011] que seleciona um conjunto de parâmetro e valores exaustivamente combinados entre si, com o objetivo de encontrar a melhor configuração para o *kernel*. Sabendo da complexidade de tal busca que cresce exponencialmente com o número de parâmetros, definiu-se um conjunto de parâmetros para cada *kernel* seguindo algumas diretrizes. Para o SVM, baseou-se no [Hsu et al. 2003], e para os demais seguiu-se o guia do usuário do Auto-WEKA [Kotthoff et al. 2017].

5.4. Serviços do T-Incident

Os resultados do modelo baseado em aprendizagem permitem compreender o melhor agrupamento espaço-temporal e o conjunto de métodos de NLP para filtragem dos textos da LBSM, e então, descrever com precisão os eventos rodoviários. Baseado nisso, produziu-se o serviço de detecção de eventos incidentes e não-incidentes e, também, o serviço de descrição dos eventos.

Uma vez identificado o evento, é possível a análise do seu contexto. Para isso, foi realizado um processo de sumarização textual, com o objetivo de criar uma versão curta e coerente que descrevesse o cenário do incidente. Aplicou-se a sumarização de texto em um grupo de *tweets* rotulados por tipo de incidente e hora e, também, pelo id do incidente. O resultado do processo fornece uma breve descrição, permitindo disponibilizar para usuários e planejadores de tráfego o ponto de vista dos usuários de LBSM em relação aos eventos de trânsito e pontos de interesse.

Neste campo, existem dois métodos de sumarização de texto, *Extrativo* e *Abstrativo*. O primeiro, seleciona os *tweets* classificando-os por sua relevância e escolhendo os mais importantes, de forma ordenada, para o significado do evento. Já o método abstrativo, procura gerar novas frases para capturar o significado do evento. Para esta versão do T-Incident, foi desenvolvido um serviço de descrição, utilizando o método extrativo.

6. Avaliação

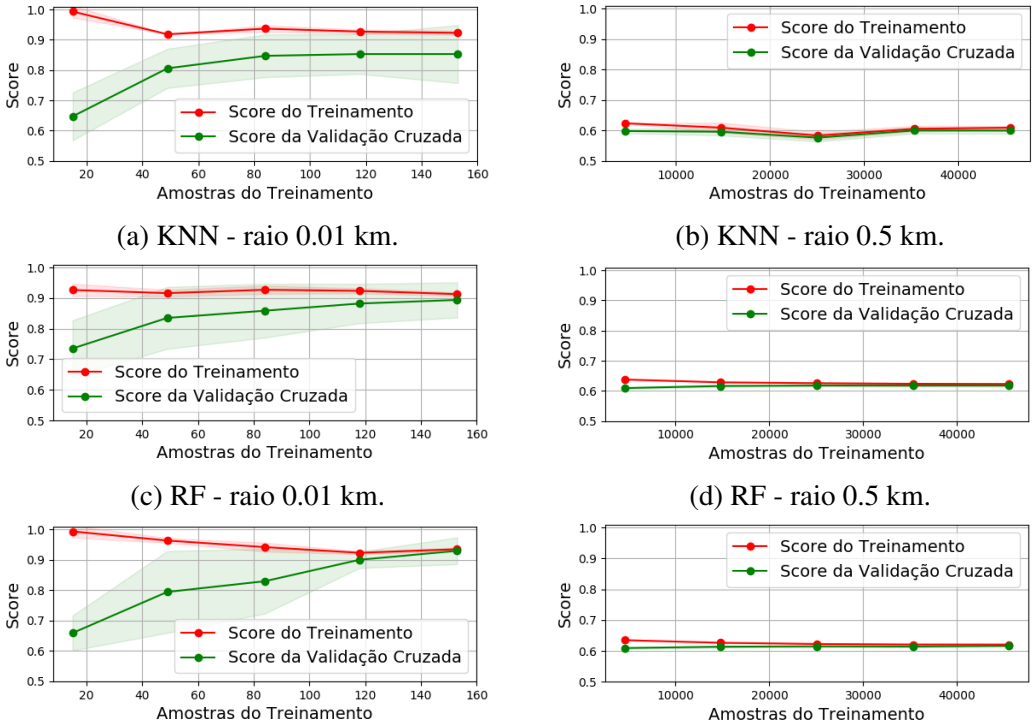
Nesta seção, descreve-se a avaliação de desempenho do T-Incident sobre um conjunto de algoritmos de classificação e modelo de agrupamento espaço-temporal como descrito na Seção 5. A seguir, apresentam-se os serviços do T-Incident para detectar e enriquecer a descrição dos eventos.

6.1. Detecção de Incidentes

A abordagem de detecção de incidentes foi baseada em uma análise exploratória de algoritmos de classificação, hiper-parâmetros e raios. Para validar os conjuntos de treinamento, foi utilizada a validação cruzada com 10 partições. A Figura 6 mostra as curvas de aprendizagem para cada *kernel* executando o modelo espaço-temporal com raio de 0,01 e 0,5 km, como exemplos. O objetivo principal é provar a generalização do modelo, procurando evitar o *over-fitting* e o *under-fitting*. Nota-se que o raio de 0,01 km (Figura 6a, 6c e 6e) fornece o melhor *score*, em torno de 90%, na maioria dos *kernels* após 140 amostras de treinamento, onde se vê as curvas convergindo e a estabilização. No entanto, poucos dados são considerados limitantes para a exploração do serviço de descrição de eventos.

Aumentando o raio, é possível observar as curvas caindo, como mostrado no maior raio na Figura 6b, 6d, e 6f. Utilizando o raio de 0,5 km, observa-se um *score* entre 58%

e 65%. Diminuindo o raio para 0,4 km, percebe-se que o *score* fica acima de 61% e abaixo 65%. Os raios entre 0,3 e 0,2 km, mostram resultados muito próximos, acima de 65% e abaixo 70%, na média. Utilizando 0,1 km, observa-se um *score* de 70%, e entre 75% e 80% considerando o raio de 0,05 km. Portanto, nota-se o *trade-off* entre maiores raios; mais dados agrupados; piores resultados e raios menores; menos dados; melhores resultado. Assim, a metodologia foi capaz de fornecer um modelo de generalização para detectar incidentes.



(a) KNN - raio 0.01 km.

(b) KNN - raio 0.5 km.

(c) RF - raio 0.01 km.

(d) RF - raio 0.5 km.

(e) SVM - raio 0.01 km.

(f) SVM - raio 0.5 km.

Figura 6: Curvas de aprendizagem de um dado *kernel* e modelo espaço-temporal.

A seguir, avaliou-se três métricas na validação cruzada e no teste. i) *F1 Score:* pondera a média entre a *Precisão* e o *Recall*. Está pontuação leva em consideração tanto os falso positivos como os falso negativos $(2 \times Recall \times Precisão)/(Recall + Precisão)$; ii) *Recall:* leva em consideração a eficiência em ser detectado os positivos $(TP/TP + FN)$; iii) *Precisão:* é a razão da observações positivas previstas corretamente para o total de obervações positivas previstas $(TP/TP + FP)$.

A Figura 7 mostra as curvas de aprendizagem para cada modelo espaço-temporal. Sendo o melhor modelo de raio 0,01 km, com *score* de Teste acima de 90% em todas as métricas avaliadas. Contudo, considera-se um bom resultado valores acima de 70%, devido a qualidade dos dados da LBSM. Desse modo, pode-se utilizar o raio de 0,1 km mantendo o *F1 sore*, *Recall* e *Precisão* em torno de 75% em média. Após esse raio de agrupamento, observa-se uma divergência e diminuição entre os *scores* de validação e teste, que pode ser explicado com o aumento da interseção dos conjunto de características dos incidentes e não-incidentes. É importante ressaltar que a generalização do modelo de aprendizado, depende da coerência entre as métricas avaliadas na validação cruzada e teste, ou seja, os resultados nesses dois contextos devem ser próximos.

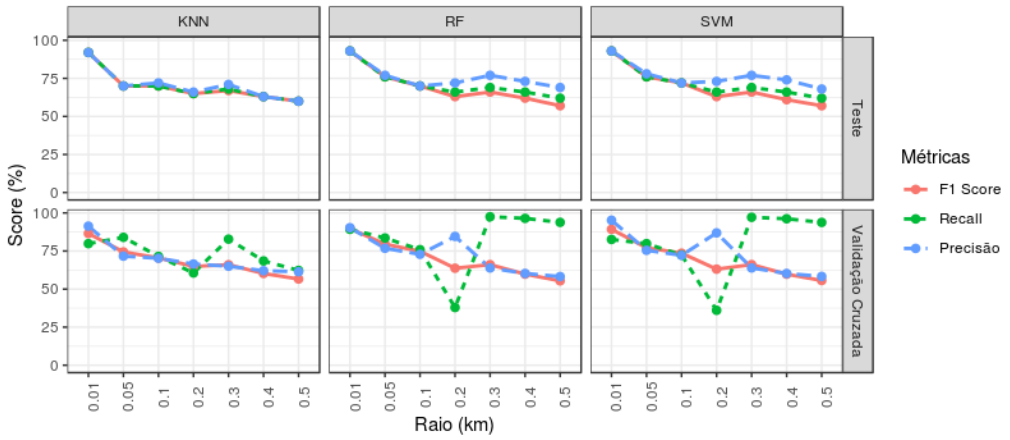


Figura 7: Resultado da classificação baseada em diferentes *kernels* e métricas.

6.2. Descrição do Evento

Os resultados observados no estágio de detecção permitiram compreender o melhor modelo de agrupamento espaço-temporal que cerca com precisão o evento, mesmo sob influência dos aspectos dos dados, como mencionado na Seção 3. Desse modo, conduziu-se o processo de sumarização, baseado no método de *Extração*, criando uma versão curta e coerente sobre o evento. Perceba que, para esta análise, foram usados raios entre 0,01 e 0,1 km, baseado no *trade-off* entre acurácia e tamanho da amostra de dados.

Como exemplo do serviço de descrição do T-Incident, utilizou-se o modelo de agrupamento espaço-temporal de 0,01 km. O texto a seguir resume um evento específico de incidente. Foram destacadas as palavras para tornar o texto mais claro. Com essa análise, o objetivo é permitir que os usuários e os administradores de estradas compreendam e decidam o que pode ser feito a respeito.

> **Cleared: Construction on #FranklinDRooseveltDrive** SB from Exit 9 - East 42nd Street to 34 street; **Updated**: Incident on #FranklinDRooseveltDrive SB at Exit 9 - East 42nd Street; **Cleared: Incident** on #FranklinDRooseveltDrive SB at Exit 9 - East 42nd Street; **Incident** on #FranklinDRooseveltDrive SB at Exit 9 - East 42nd Street; **Closure** on #FranklinDRooseveltDrive NB at Exit 9 - East 42nd Street; **Cleared**: Closure on #FranklinDRooseveltDrive NB at Exit 9 - East 42nd Street; **Construction** on #FranklinDRooseveltDrive Both directions at Exit 9 - East 42nd Street

Ao mesmo tempo, utilizando o agrupamento espaço-temporal com raio de 0,1 km, por exemplo, analisou-se um evento não-incidente específico no *Town Hall* e suas vizinhança. O texto abaixo resume a área, evidenciando as tendências de lugares extraídos através da impressão de usuários. Dessa forma, é possível descobrir os lugares culturais, onde reservar um quarto de hotel e até mesmo onde almoçar.

> **Open House New York** Sunday Stop 1! **Town Hall**. It was never taken over by the Broadway **theatre** giants because ther..; #30DaysForMyArt DAY 16: "Go see a **broadway show**."It's simple: There is NOTHING like a broadway show. I've lived in; **Beastie Boys Book**: Live; Direct with Adam Horovitz; Michael Diamond: The **Town Hall**; Good morning **Times Square**. Bad I have to leave today! (**@Millennium Broadway Hotel** - @millenniumpr in New York, NY); Good night! (**@Millennium Broadway Hotel** - @millenniumpr in New York, NY); YEP! I Like Wrestling Podcast #45: WWE Super Show-Down Predictions, Raw; Smack; Show Time! (@Beautiful: The Carole King Musical in New York, NY); **Mooch's book party**. Really. (@Hunt; **Fish Club** in New York, NY); Head over heels wPeppermint!!! (@Hudson Theatre - @hudsonbway for Head Over Heels in New York, NY); I had the heirloom tomato lobster salad. **Kristine had the burger** (@Burger; Lobster in New York, NY);

7. Conclusão

Neste trabalho, apresentou-se o T-Incident, uma arquitetura robusta de detecção e enriquecimento de eventos rodoviários baseado em fusão de dados heterogêneos do ITS. Para

isso, foi desenvolvido uma modelo de agrupamento espaço-temporal para fundir dados de incidentes, não-incidentes e dados de LBSM. Em seguida, focou-se na detecção de padrões nas mídias sociais para descrever o evento de incidente e suas vizinhanças. Utilizando um modelo baseado em aprendizagem e técnicas de NLP, foi possível fornecer serviços de detecção e descrição de eventos.

Os resultados mostraram os melhores conjuntos de parâmetros que podem ser utilizados no T-Incident, permitindo fornecer a detecção de incidentes, com *score* de até 90% de *F1 score*, *Recall* e *Precisão*. Além de um serviço de descrição de eventos que permite entregar aos usuários e planejadores de tráfego o ponto de vista dos usuários em relação aos eventos de trânsito e pontos de interesse. Como trabalhos futuros, pretende-se estender o T-Incident adicionando mais contas especialistas, melhorando a identificação atual e a descrição dos eventos. Baseado nos resultados do T-Incident, pretende-se também avançar na previsão de incidentes e avaliação do tempo de duração do incidente.

Referências

Bazzan, A. L. and Klügl, F. (2013). *Introduction to intelligent systems in traffic and transportation*. Morgan & Claypool Publishers.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Jockers, M. (2017). syuzhet: Extracts sentiment and sentiment-derived plot arcs from text.

Khaleghi, B., Khamis, A., Karray, F., and Razavi, S. (2013a). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*.

Khaleghi, B., Khamis, A., Karray, F. O., and Razavi, S. N. (2013b). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44.

Kotthoff, L., Thornton, C., and Hutter, F. (2017). User guide for auto-weka version 2.6. *Dept. Comput. Sci., Univ. British Columbia, BETA lab, Vancouver, BC, Canada, Tech. Rep*, 2.

Nguyen, H., Liu, W., Rivera, P., and Chen, F. (2016). Trafficwatch: Real-time traffic incident detection and monitoring using social media. In *PAKDD*, pages 540–551. Springer.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of MLR*, 12:2825–2830.

Pereira, F. C., Rodrigues, F., and Ben-Akiva, M. (2013). Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, 37:177–192.

Rettore, P. H., Santos, B. P., Campolina, A. B., Villas, L. A., and Loureiro, A. A. (2016). Towards Intra-Vehicular Sensor Data Fusion. *19th International Conference on ITS*.

Santos, B. P., Rettore, P. H., Ramos, H. S., Vieira, L. F. M., and A.F. Loureiro, A. (2018). Enriching traffic information with a spatiotemporal model based on social media. In *ISCC*, Natal, Brazil.

Xu, S., Li, S., and Wen, R. (2018). Sensing and detecting traffic events using geosocial media data: A review. *Computers, Environment and Urban Systems*, (June).

Yazici, M. A., Mudigonda, S., and Kamga, C. (2017). Incident detection through twitter: Organization versus personal accounts. *TRR: Journal of the TRB*, (2643):121–128.

Zhang, Z., He, Q., Gao, J., and Ni, M. (2018). A deep learning approach for detecting traffic accidents from social media data. *Transportation research part C: emerging technologies*, 86:580–596.