# MÉTODOS PARA AVALIAÇÃO DE CONDIÇÕES DE TRÁFEGO A PARTIR DE DADOS DO GOOGLE TRAFFIC E DO TWITTER

**Conference Paper** · November 2018

CITATIONS
0

READS
106

**2 authors:**

Andre Borgato Morelli
University of São Paulo
**6** PUBLICATIONS **7** CITATIONS
SEE PROFILE

André Luiz Cunha
University of São Paulo
**53** PUBLICATIONS **115** CITATIONS
SEE PROFILE

# MÉTODOS PARA AVALIAÇÃO DE CONDIÇÕES DE TRÁFEGO A PARTIR DE DADOS DO GOOGLE TRAFFIC E DO TWITTER

**André Borgato Morelli**
**André Luiz Cunha**
Universidade de São Paulo
Escola de Engenharia de São Carlos

**RESUMO**
Tendo em vista a dificuldade na determinação do nível de serviço em regiões urbanas complexas, o presente trabalho propõe uma análise exploratória de dois métodos distintos para avalição de qualidade de transporte para veículos automotores em regiões urbanas. O primeiro método baseou-se na medição do comprimento de congestionamentos e lentidão na rede viária a partir de técnicas de processamento de imagens em mapas de tráfego. Para tanto foram obtidas imagens da camada de tráfego do *Google Maps* (*Google Traffic*), para intervalos de 15 minutos, na cidade de São Paulo, visando a determinação do comprimento total de vias congestionadas ou com lentidão a partir de contagem de pixels. Já o segundo método baseou-se em dados obtidos em uma rede social digital, analisados com técnicas de processamento de linguagem natural e aprendizado de máquina para obtenção da qualidade do ponto de vista do usuário. Para tanto, foram coletados comentários no *Twitter* (*tweets*) com palavras-chave relacionadas ao tráfego e classificados como positivos, negativos ou não relacionados. Durante os cinco dias úteis analisados, o volume dos comentários negativos foi comparado ao comprimento total de vias congestionadas ou com lentidão na região urbana. Os resultados apontam forte correlação entre os dois métodos, e sinalizam a possibilidade de ambos serem alternativas viáveis para a medida de qualidade de tráfego.

**ABSTRACT**
Considering the difficulty in determining the level of service in complex urban regions, this paper proposes an exploratory analysis of two different methods to evaluate quality of transport for motorized vehicles in urban regions. The first method was based on measuring the length of congestion and slow traffic in the road network, applying image processing techniques in traffic maps. For this purpose, images from the traffic layer of Google Maps (Google Traffic) were obtained for 15-minute intervals in the city of São Paulo to determine the total length of congested or slow paths through pixel counting. The second method was based on data obtained in a digital social network, analyzed using natural language processing techniques and machine learning, to obtain the user's point of view on quality. For this purpose, Twitter comments (tweets) with keywords related to traffic were collected and classified as positive, negative or unrelated. During the five working days analyzed, the volume of negative comments was compared to the total length of congested or slow roads in the urban region. The results indicate a strong correlation between the two methods and points to the possibility of both being viable alternatives for measuring quality of traffic.



## 1. INTRODUÇÃO

O estudo da qualidade de sistemas de transporte é essencial ao desenvolvimento de projetos bem-sucedidos em satisfazer as necessidades dos usuários. No aspecto urbano, a avaliação dos sistemas viários para automóveis é feita, em geral, segundo as recomendações do *Highway Capacity Manual* (HCM) que preconiza a categorização das vias em seis níveis de serviço a partir da velocidade média do tráfego e razão volume por capacidade ao longo da via (TRB, 2010). Apesar de essa metodologia refletir relativamente bem a percepção de qualidade pelos usuários, ela fornece informações sobre vias isoladas (TRB, 2010) e não sobre o sistema viário urbano como um todo ou sobre como a percepção do usuário varia regionalmente. Além disso, a obtenção de medidas de velocidade média de tráfego ao longo de um sistema complexo, como o de uma grande metrópole, pode se provar difícil e custosa, tornando bem-vindas alternativas que sejam capazes de medir as condições do sistema de tráfego como um todo.

Uma fonte de dados que vem sendo explorada de maneira crescente para obtenção de informação são as redes sociais digitais, em que os usuários expressam de maneira pública seus sentimentos com relação aos mais variados temas. Segundo Gal-Tzur *et al.* (2014), mensagens

geradas por usuários têm um papel importante em diversas áreas da sociedade atual como entretenimento, política e negócios, assim, seria razoável concluir que o setor de transportes também poderia ser incluído nesse grupo. A mineração da opinião no *Twitter*, uma plataforma de *microblogging* em que os usuários podem publicar comentários de até 280 caracteres, provou-se particularmente promissora nesse aspecto, devido aos seus comentários curtos e densos em termos de informação, além da plataforma disponibilizar gratuitamente seu banco de dados de comentários públicos. Nesse contexto, a plataforma se mostra muito promissora como fonte de mineração de dados, podendo oferecer informações que possam qualificar os sistemas de transporte a partir do ponto de vista do usuário. Com este tipo de abordagem, pode ser possível detectar diferentes tolerâncias ao tráfego em diferentes regiões do Brasil devido a fatores socioculturais, diferente do que é proposto pelo modelo do HCM. Além disto, obter a opinião dos usuários em grandes escalas, como a nível de cidade, pode oferecer uma oportunidade para detectar as regiões urbanas mais bem sucedidas em sua gestão de tráfego e auxiliar em possíveis pesquisas que visem relacionar políticas de trânsito, forma urbana, conectividade da rede, ocupação do solo e outros fatores tipicamente abordados em planos diretores com a qualidade de fato percebida pelos usuários nas cidades.

Outra fonte interessante são os dados de sensoriamento remoto de dispositivos móveis, que são cedidos por grandes volumes de usuários a empresas como *Google* e *TomTom*. A partir da localização dos usuários em tempo real, essas empresas são capazes de gerar estimativas de velocidade média de tráfego e tempo médio de trajeto, permitindo avaliar congestionamentos e lentidão, que podem ser parâmetros viáveis para a caracterização do tráfego de uma grande cidade com relação a seu tráfego em escalas maiores que a proposta pelo HCM.

Tendo isso em mente, e considerando-se a escassez de trabalhos que relacionem opiniões de redes sociais com o comprimento de congestionamento em grandes cidades, o objetivo deste artigo é fazer uma análise exploratória acerca da negatividade dos comentários dos usuários no *Twitter* e do congestionamento obtido por meio de mapeamento do tráfego na cidade de São Paulo, investigando a correlação entre os resultados encontrados nas duas plataformas. Este trabalho faz parte de uma pesquisa mais ampla que visa relacionar fatores de forma e conectividade de redes ao nível de serviço de sistemas urbanos de tráfego, atualmente em desenvolvimento em uma dissertação de mestrado.

## 2. TRABALHOS ANTERIORES

Desde seu surgimento, as redes sociais digitais cresceram em termos de influência e alcance da população global, acumulando nesse processo um volume crescente de informação. Neste contexto, trabalhos recentes tentaram extrair dados dessas plataformas que pudessem ajudar em tomadas de decisões nas mais variadas vertentes de sistemas transportes. Alguns dos trabalhos mais comuns na área são relacionados à detecção automática de incidentes relacionados ao tráfego como acidentes e bloqueios a partir de comentários de redes sociais (Mai e Hranac, 2013; Schulz *et al,* 2013). Também são comuns trabalhos voltados diretamente à inferência das condições de trânsito a partir desse tipo de dado, alguns utilizando contas oficiais de agências governamentais ou redes jornalísticas como fontes, o que exclui a necessidade de tratamento dos dados para determinar se os comentários são ou não relevantes (Albuquerque *et al*, 2012; Oliveira *et al*, 2013; Pathak *et al*, 2015) e outros utilizando dados georreferenciados que, apesar de pouco numerosos, contém informações precisas acerca da localidade das ocorrências (Pan *et al,* 2013; Wang *et al*, 2015; Gong *et al,* 2015). Existe, contudo, um volume de informação ainda pouco explorado nas redes sociais para fins de transportes: os dados não

georreferenciados. Segundo Mai e Hranac (2013), em torno de 1,3% dos *tweets* possuem informação sobre localidade do evento, o que dificulta a obtenção de informação precisa acerca de onde os congestionamentos ocorrem. Assim, a análise conduzida com dados georreferenciados implica a exclusão da maior parte dos tweets que possivelmente seriam relevantes à análise em troca da maior precisão de localização do ocorrido.

É notável também o crescimento do número de pesquisas utilizando dados de dispositivos embarcados em veículos automotores, ou relativos a aparelhos celulares para auxiliar na inferência de parâmetros de tráfego. A premissa básica para esse tipo de abordagem é que monitorar os deslocamentos de usuários através de tecnologia GPS pode fornecer uma boa estimativa da velocidade média do tráfego da região em que o dispositivo se situa ou capturar comportamentos de grandes grupos de motoristas. Nesse contexto, Amin *et al.* (2008) conduziram um experimento em uma rodovia na Califórnia, EUA a fim de estimar tempo de percurso e velocidade média na via a partir de dados do GPS de 100 celulares Nokia e compará-la ao obtido por meio de sensores fixos dispostos na rodovia. Os autores chegaram à conclusão que é possível estimar a velocidade de fluxo e tempo de viagem em uma via com uma penetração de menos de 5% de veículos com tecnologia embarcada. Mais recentemente, quando a plataforma *Google Maps* disponibilizou sua *Application Programming Interface* (API) para acesso aos dados de tráfego de seus milhares de usuários, Wang e Yanqing (2011) foram capazes de estimar uma matriz Origem-Destino a partir de dados de tráfego provindos da API do *Google Maps*. Os dados obtidos foram relativos à velocidade média de trafego em vias, que é obtido a partir do sensoriamento remoto de usuários da plataforma. Existe, contudo, o problema de a API possuir restrições com relação à quantidade de informação que podia ser obtida de maneira gratuita, como os próprios autores ressaltam em seu trabalho.

## 3. MÉTODO PROPOSTO

*Quantificação da opinião do usuário:* No primeiro método, a qualidade é considerada pelo ponto de vista do usuário a partir de comentários do *Twitter* na cidade de São Paulo extraídos a partir de palavras chave relacionadas ao transporte. Esses comentários são processados com técnicas de Processamento de Linguagem Natural (PLN) e classificados como positivos, negativos ou não relacionados a partir de algoritmos de aprendizado de máquina, para determinação das frequências relativas horárias de opiniões negativas sobre o transporte.

*Estimativa do congestionamento na rede viária:* Já o segundo método, foi baseado no comprimento de vias congestionadas ou com lentidão como um indicador geral da qualidade do tráfego. Esse parâmetro foi obtido através do processamento de imagens de mapas de tráfego obtido da plataforma *Google Maps*.

*Correlação entre opinião e congestionamento:* Os dois métodos anteriores podem ser comparados para investigar a relação entre o comprimento total de congestionamentos ou lentidão no sistema e a qualidade como percebida pelos usuários do *Twitter*. O fluxograma da Figura 1 mostra a sequência geral de análise e as etapas do processo são explicadas nos tópicos a seguir.

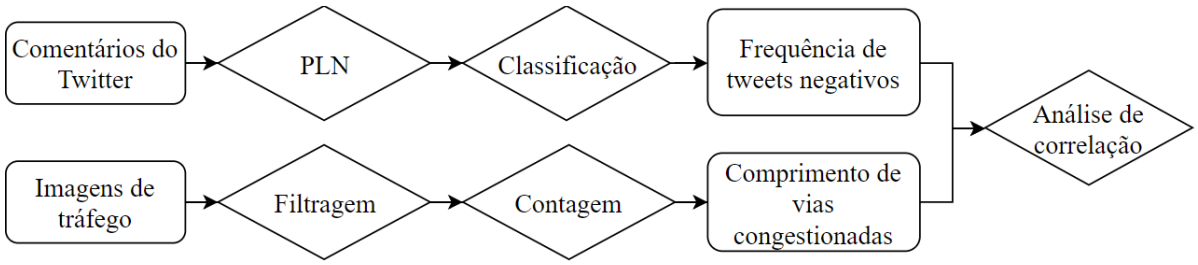


**Figura 1:** Etapas de desenvolvimento e comparação de métodos.

### 3.1. Avaliação da opinião dos usuários

*3.1.1. Extração de dados do* Twitter

A extração de dados foi conduzida utilizando a biblioteca para a linguagem *Python* de programação, *tweepy* (Roesslein, 2009). O *Twitter* permite que sejam buscados gratuitamente *tweets* públicos, dos últimos sete dias, por meio de uma ou mais palavras-chave. Assim, neste trabalho foram utilizadas as palavras chave: "trânsito", "tráfego" e "congestionamento" com adição de algumas variações como as palavras com ausência de acentuação como "transito" e possíveis erros comuns como "congetionamento". Além disso, foram filtrados *retweets,* ou seja, *tweets* de outras pessoas que são republicados. Essas palavras-chave foram escolhidas por estarem claramente relacionadas ao tráfego e facilitarem uma análise com restrição de banco de dados para cinco dias de tráfego.

Para fins de comparação com os dados obtidos dos mapas de tráfego, os *tweets* também foram coletados de 4 a 8 de junho de 2018. Além disso, o banco de dados foi filtrado para compreender apenas os tweets dos usuários do Twitter que declaram viver na cidade de São Paulo.

*3.1.2. Processamento de linguagem natural*

Um processo comum empregado para facilitar a tarefa de classificação é a normalização e tokenização de texto em linguagem natural. Normalização de texto é o processo de identificação de números, abreviações, acrônimos e anomalias idiomáticas com subsequente transformação desses em suas formas mais informativas, a fim de conduzir uma análise mais abrangente dos dados. Algumas tarefas simples neste aspecto, que foram aplicadas ao banco de dados textual do *Twitter*, são as de redução de texto, como a transformação de letras maiúsculas em minúsculas, remoção de caracteres especiais, que não expressam significado determinado, ou, em alguns casos, remoção de pontuação. Outras técnicas um pouco mais complexas envolvem a separação do texto em diferentes níveis como palavras, sentenças e tópicos. Estes algoritmos são chamados algoritmos de tokenização (Bird *et al.*, 2009). Para este artigo, foi utilizada a tokenização de palavras em três níveis diferentes: palavras individualmente, palavras com seu vizinho à frente e palavras com seus dois vizinhos à frente, de forma que a frase "engarrafamento na avenida sete" é decomposta nos *tokens*:

1- "engarrafamento", "na", "avenida", "sete";
2- "engarrafamento na", "na avenida", "avenida sete";
3- "engarrafamento na avenida", "na avenida sete".

No caso específico de comentários retirados do *Twitter*, considerou-se vantajosa a eliminação de referências a outras contas do *Twitter*, caracterizada na plataforma pelo sinal "@" seguido do nome da conta à qual a referência é direcionada. Além disso, outros tipos de referência externa podem não acrescentar informação relevante à análise, como links externos que também

foram removidos na análise conduzida.

Outra estratégia empregada neste trabalho foi a marcação de ênfase. Duas palavras “adoro” e “adooooro” têm significados similares, cuja única diferença é a ênfase, mas no método de vetorização empregado seriam consideradas palavras completamente diferentes. Além disso, palavras com pequenas variações como “adoooro” e “adooooro” também seriam colocadas em categorias diferentes, mesmo que ambas tenham mesmo significado e ênfase. Neste caso, uma palavra deste tipo é normalizada para sua forma mais simples e marcada com um marcador de ênfase (&). A Tabela 1 exemplifica casos de marcação de ênfase comuns levados em consideração nesse trabalho. O marcador de ênfase escolhido foi o símbolo “&” devido à sua raridade em vocabulário português, diferenciando as palavras enfatizadas das comuns.

**Tabela 1:** Exemplos de marcação de ênfase.

| Frase original | Frase normalizada | *Tokens* de uma palavra |
|---|---|---|
| "Que Trânsito" | "que transito" | "que", "transito" |
| "Que TRÂNSITO" | "que transito&" | "que", "transito&" |
| "Que trâaaansito" | "que transito&" | "que", "transito&" |
| "Que TRÂAAAANSITO" | "que transito&&" | "que", "transito&&" |
| "Que transito!!!!!" | "que transito!&" | "que", "transito", "!&" |

Posteriormente, a fim de aplicar métodos de classificação ao banco de dados, os *tokens* obtidos do texto foram transformados em vetores. Para tanto, foi utilizada a técnica *bag-of-words,* que consiste na contagem dos *tokens* presentes em uma frase. Em um banco de dados vetorizado, cada linha refere-se a um tweet e cada coluna refere-se à contagem de um *token* possível entre um conjunto predefinido contendo, no caso deste trabalho, os 6.000 *tokens* mais frequentes no banco de dados utilizado. A Tabela 2 mostra um exemplo deste tipo de representação do banco de dados.

**Tabela 2:** Exemplo de *tweets* vetorizados com o método *bag-of-words.*

| Tweet | horrível | está | ... | Trânsito | congestionamento |
|---|---|---|---|---|---|
| "Esse trânsito está horrível" | 1 | 1 | ... | 1 | 0 |
| "Nossa, que congestionamento" | 0 | 0 | ... | 0 | 1 |
| "Trânsito horrível, muito horrível" | 2 | 0 | ... | 1 | 0 |

*3.1.3. Classificadores*

Foram utilizados três tipos de classificador: regressão logística, *Suport Vector Machines* (SVM) e *Naive Bayes*. Os dois primeiros são capazes de fazer distinção entre apenas duas categorias, mas, como neste trabalho foi feita a classificação de três categorias, positivo, negativo e não relacionado, os classificadores foram empregados com um método composto, a classificação um-contra-todos. A avaliação dos resultados foi feita utilizando as matrizes de contingência dos classificadores e algumas métricas de desempenho tipicamente utilizadas. Segundo Goutte e Gaussier (2005), as métricas unidimensionais mais frequentemente utilizadas para avaliação de modelos são precisão, *recall* e F1. Neste artigo foi utilizada, além das citadas, a exatidão como parâmetro de avaliação, já que essa é uma medida mais geral do modelo um-contra-todos.

*Precisão:*

A precisão é definida como o número de verdadeiros positivos sobre a soma de verdadeiros

positivos (VP) e falsos positivos (FP), como:

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \quad (1)$$

A precisão é uma medida que descreve a eficiência do classificador em termos de classificar eventos positivos. Uma precisão alta significa que quando o algoritmo classifica um evento como positivo existe grande chance de que ele seja de fato positivo. Contudo, isso não implica uma boa taxa de acerto já que um classificador conservador pode classificar como verdadeiros positivos apenas os casos em que a probabilidade de o evento ser positivo for muito alta, resultando em um modelo com grande taxa de falsos negativos.

*Recall:*
O *recall* é também conhecido como a taxa de verdadeiros positivos, ou seja, o número de verdadeiros positivos (VP) sobre a soma de verdadeiros positivos e falsos negativos (FN), como:

$$\text{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad (2)$$

O *recall* é uma medida que descreve a eficiência do classificador em detectar eventos positivos. Assim, um classificador com alto *recall* conseguirá encontrar quase todos os eventos positivos em uma amostra. Existe, contudo, o risco de um classificador do tipo ser pouco conservador, gerando uma grande taxa de falsos positivos.

*Fator F1:*
Em geral, um classificador pode ter alta precisão, mas perder eficiência no *recall* ou vice-versa. O fator F1 é uma medida que busca relacionar precisão e *recall* de forma a capturar o comportamento médio. Assim, o fator F1 é definido como a média harmônica entre precisão e *recall,* como:

$$\text{F1} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \quad (3)$$

*Exatidão:*
A exatidão mede o desempenho global do classificador em fazer previsões corretas e leva em consideração tanto os verdadeiros positivos quanto os verdadeiros negativos em relação ao total de previsões, como:

$$\text{Exatidão} = \frac{\text{VP+VN}}{\text{VP+VN+FP+FN}} \quad (4)$$

**3.2. Estimativa de congestionamento em rede**
O monitoramento do tráfego da cidade foi conduzido a partir do processamento de imagens de mapas da camada de tráfego do *Google Maps*, obtidos a cada 15 minutos durante o período das 0h00min do dia 4 de junho até às 23h45min do dia 8 de junho. Esse intervalo foi escolhido por não apresentar nenhuma ocorrência significativa como dias de fim de semana, feriados, paralizações ou greves, possibilitando aos autores tratar os dados como dados típicos. A coleta foi feita por meio da captura de tela da camada de tráfego do *Google Maps* em pequenas regiões da cidade com posterior imersão das imagens a partir de suas coordenadas para a formação do mapa completo da cidade. O intervalo de 15 minutos foi escolhido por ser o menor tempo em que a captura e processamento das imagens pudesse ocorrer no computador utilizado. Os mapas

de tráfego foram analisados em um retângulo projetado sobre as coordenadas da cidade de São Paulo, no Nível de Zoom (z) 15 do *Google Maps*, que permitiu obter o tráfego de todas as vias em uma região. Níveis de zoom maiores não adicionariam informação e aumentariam o tempo de processamento, enquanto níveis menores causariam perda da informação de vias menos utilizadas. As extremidades do mapa geral da cidade são os pontos:

- Canto superior esquerdo: (-23,46945° Lat, -46,78467° Lon)
- Canto inferior direito: (-23,63756° Lat, -46,52254° Lon)

Todos as vias contidas no intervalo descrito acima foram consideradas na análise. O processo de mensuração processo se inicia com uma filtragem dos pixels de interesse (vermelhos e vinho) da imagem a partir de suas cores, como pode ser observado na Figura 2.

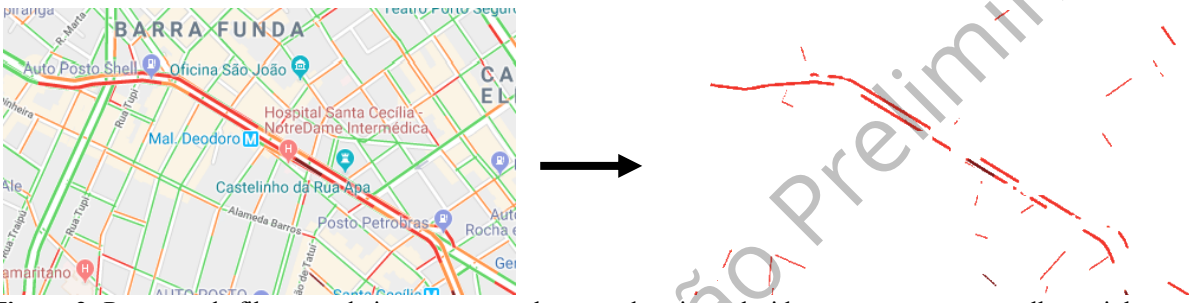


**Figura 2:** Processo de filtragem de imagens para obtenção das vias coloridas com as cores vermelho e vinho. Imagens obtidas da plataforma *Google Maps* (https://www.google.com.br/maps)

O comprimento total das vias foi determinado medindo-se o comprimento de um alinhamento na imagem, em pixels, e posteriormente convertendo essa medida para quilômetros. Essa conversão é possível através da determinação de um fator médio de conversão dado em km/pixel, obtido utilizando-se as equações da projeção *Google Web Mercator*, utilizada para gerar os mapas na plataforma (Battersby *et al*, 2014). Nessa projeção é possível calcular uma coordenada, em pixels como:

$$x = \frac{256}{2\pi} \cdot 2^z(\lambda+\pi) \tag{5}$$

$$y = \frac{256}{2\pi} \cdot 2^z \left(\pi - \ln\left[\tan\left(\frac{\pi}{4}+\frac{\phi}{2}\right)\right]\right) \tag{6}$$

em que $x$: Coordenada horizontal do ponto no mapa projetado [pixels];
$y$: Coordenada vertical do ponto no mapa projetado [pixels];
$z$: Nível de Zoom do mapa do *Google Maps*;
$\lambda$: Latitude do ponto [°];
$\phi$: Longitude do ponto [°].

Sabendo-se, portanto, a coordenada de dois pontos no mapa é possível calcular a distância euclidiana entre esses pontos, em pixels, a partir das coordenadas calculadas com as Equações 5 e 6. A partir disso, o fator de conversão é calculado por:

$$\rho_m = \frac{D_R}{D_I} \tag{7}$$

em que $\rho_m$: Fator médio de conversão para distâncias em pixels [km/pixel];
$D_R$: Distância real entre os pontos [km];

$D_I$: Distância entre os pontos na imagem [pixels].

Deve-se ressaltar que se trata de um fator médio que é válido apenas para regiões relativamente pequenas do globo terrestre e pode ser utilizado para uma cidade quando não é necessária grande precisão nas medidas, como no caso deste trabalho. O valor calculado para a diagonal que passa pelos pontos do canto superior esquerdo e inferior direito do mapa utilizado da cidade de São Paulo é de 4,374 x $10^{-3}$ km/pixel, contudo, dada a distorção que ocorre na projeção de Mercator, quanto maior a latitude de uma região, menor deverá ser esse fator.

### 3.3. Correlação entre opinião e congestionamento

Os métodos propostos para a avaliação da qualidade do tráfego se baseiam em métricas diferentes, contudo ambos buscam estimar a insatisfação dos motoristas com o tráfego. O comprimento total de vias congestionadas ou com lentidão é um parâmetro geral de tráfego que busca capturar a dificuldade de um motorista em percorrer uma rede viária. Por outro lado, a análise dos comentários do *Twitter* busca diretamente a opinião do usuário do sistema, já que o volume de comentários negativos na rede social é um bom indicador do impacto do tráfego na vida do usuário e como isso varia ao longo do tempo. Neste caso, busca-se relacionar essas duas medidas através de análise de correlação e verificar o impacto do carregamento do tráfego na opinião pública expressada em redes sociais.

## 4. RESULTADOS

### 4.1. Classificação de *tweets*

O treinamento dos três tipos de classificador utilizou um banco com 2.034 *tweets* manualmente classificados em três categorias: “Comentário Negativo”, “Comentário Positivo” e “Comentário não Relacionado ao Trânsito”. Para testar o desempenho do classificador, os dados foram divididos em um grupo de treinamento com 80% das entradas e um grupo de teste com o restante. Além disso, com objetivo de evitar incorrer em viés de classificação, o banco de dados foi balanceado para as três categorias, de forma que tanto o treinamento como o teste ocorrem com entradas das três categorias em números iguais. Os textos dos *tweets* foram normalizados com:

- Substituição de maiúsculas;
- Eliminação de referências de contas do *Twitter* como em “@NomeDoUsuario”;
- Marcação de ênfase para uso de letras maiúsculas no meio de frases como em “TRÂNSITO NA AVENIDA”;
- Marcação de ênfase para repetição de caracteres em excesso;
- Remoção de acentos de palavras;
- Marcação de números.

A Tabela 3 mostra as medidas básicas de desempenho, enquanto a Figura 3 contém as matrizes de contingência de cada classificador. Nota-se que o classificador SVM obteve a melhor desempenho segundo a métrica F1 em todas as categorias, além de também ter obtido a melhor exatidão global. Joachims (1998), argumenta que classificadores SVM se comportam bem para tarefas de classificação de texto devido, entre outros fatores, a possuírem proteção contra *overfitting*, caraterística fundamental para sistemas com milhares de dimensões a serem avaliadas. Desta forma, o classificador SVM se mostrou mais vantajoso.

**Tabela 3:** Medidas de desempenho dos classificadores.

| Modelo | Regressão logística | | | *Naive Bayes* | | | SVM | | |
|---|---|---|---|---|---|---|---|---|---|
| Categoria | -1 | 0 | 1 | -1 | 0 | 1 | -1 | 0 | 1 |
| Precisão | 87% | 83% | 81% | 70% | 91% | 85% | 87% | 89% | 85% |
| Recall | 75% | 89% | 91% | 83% | 83% | 82% | 80% | 95% | 87% |
| F1 | 0.80 | 0.86 | 0.86 | 0.76 | 0.87 | 0.84 | 0.84 | 0.92 | 0.86 |
| Exatidão | | 84% | | | 82% | | | 87% | |

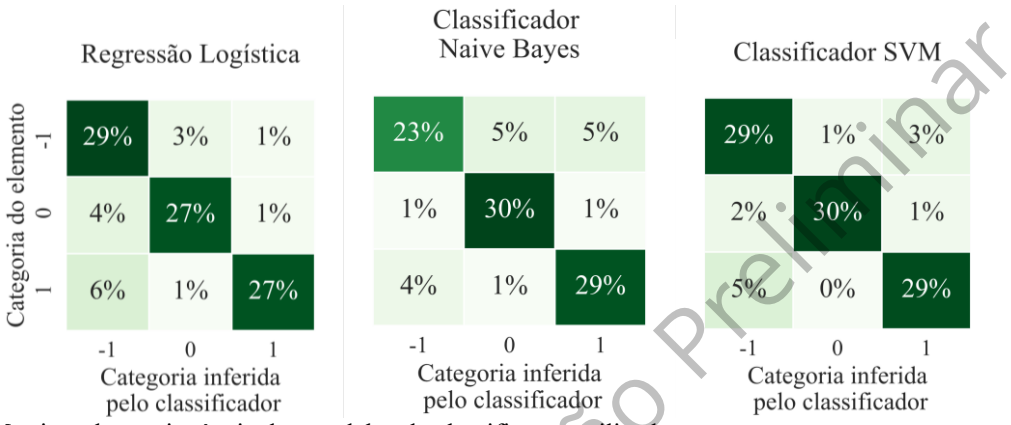


**Figura 3:** Matrizes de contingência dos modelos de classificação utilizados.

## 4.2. Verificação da distribuição de *tweets*

Com os *tweets* do período e o classificador definido, foi possível estabelecer a distribuição média do volume de *tweets* negativos por hora do dia para essa semana, como mostrado no gráfico da Figura 4. É possível notar que existem dois picos bem definidos, um por volta das 9h e outro por volta das 19h.

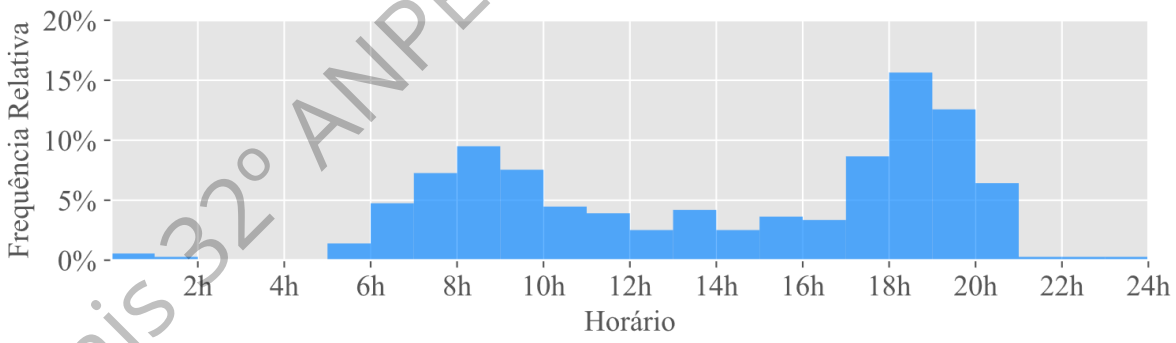


**Figura 4:** Histograma de *tweets* negativos referentes ao tráfego da cidade de São Paulo. Medidas médias para o período de 4 a 8 de junho de 2018.

## 4.3. Obtenção do comprimento de vias congestionadas

Como exposto anteriormente, os mapas de tráfego foram obtidos da plataforma a cada 15 minutos no período avaliado. Através dos processos de filtragem de pixels de interesse, contagem linear de pixels e conversão para distâncias reais, foi então obtido o comprimento total de vias em função do tempo. Na Figura 5 pode ser observado o comportamento médio do congestionamento e lentidão na cidade para o período avaliado.

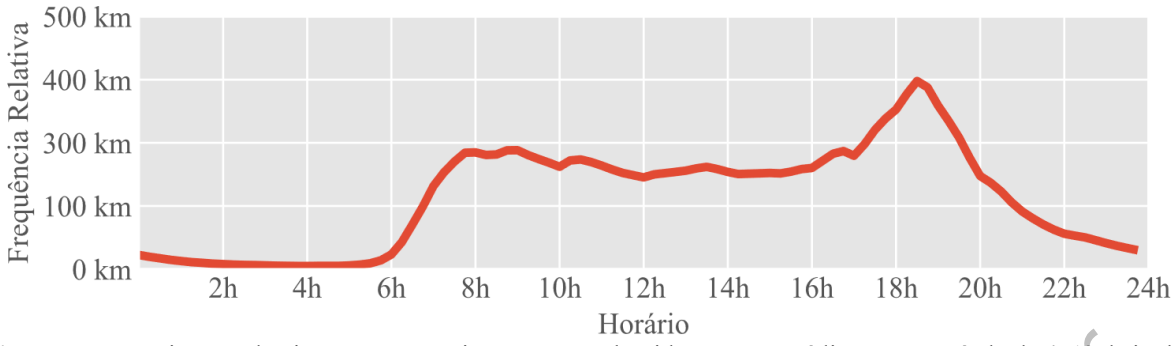


**Figura 5:** Comprimento de vias com congestionamento ou lentidão. Curva média para o período de 4 a 8 de junho de 2018

### 4.2. Relação entre *tweets* e o congestionamento

Pode-se observar o comportamento médio do tráfego na Figura 6, bem como a frequência relativa média dos *tweets*. A opinião negativa dos usuários, como medida no *Twitter*, em geral é mais alta para períodos em que existe mais lentidão no tráfego, particularmente nos picos da manhã e da tarde. Esse resultado indica que a percepção da qualidade dos usuários está relacionada ao comprimento total de lentidão e congestionamentos.

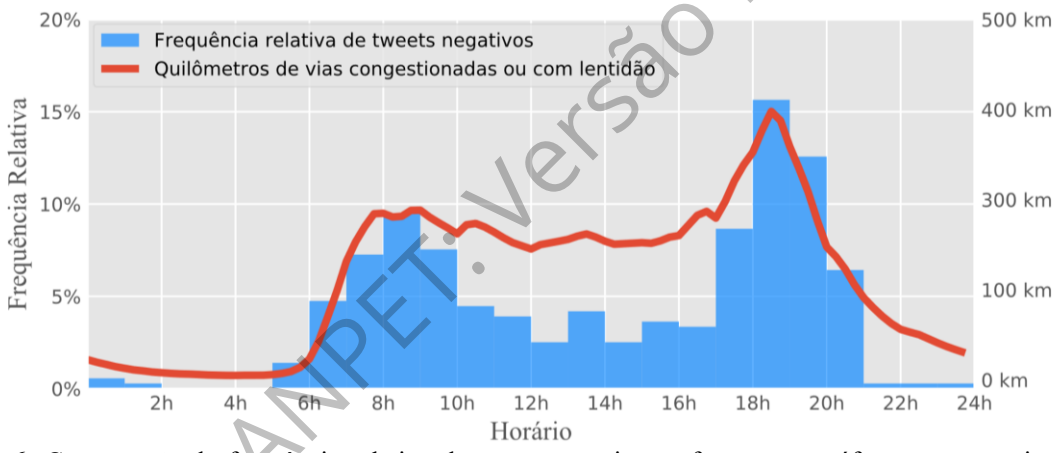


**Figura 6:** Comparação da frequência relativa de *tweets* negativos referentes ao tráfego e o comprimento de congestionamento ou lentidão na cidade de São Paulo.

Além disso, foi conduzida uma análise de correlação entre o carregamento do tráfego e a atividade do *Twitter (*Figura 7*)*. Nesta abordagem, para cada período de 30 minutos, verificou-se a condição média de tráfego e a contagem média de *tweets* durante a semana.

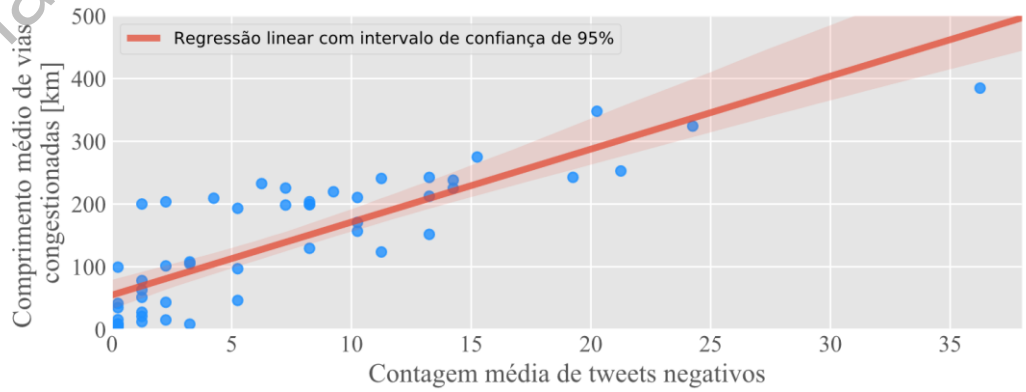


**Figura 7:** Relação linear entre comprimento médio de vias congestionadas e contagem média de *tweets* negativos

A correlação encontrada para esta análise foi de 0,84. Nota-se também uma dispersão maior para contagens baixas de *tweets*. Isso pode acontecer devido ao ruído elevado nessas faixas que se deve ao fato de a contagem de *tweets* ser, por natureza, uma variável discreta, de forma que está sujeita a grandes variações aleatórias para faixas menores que 10, nas quais um tweet a mais ou a menos pode resultar uma variação superior a 10%. Como a análise aqui mostrada é referente à média dessas contagens, o ruído tende a se dissolver com o aumento do período analisado e, neste caso, espera-se que essa dispersão seja reduzida com a expansão do período analisado para mais de uma semana.

## 5. CONSIDERAÇÕES FINAIS

Este trabalho realizou uma análise exploratória para obtenção da qualidade do tráfego na cidade de São Paulo. Para tal foram empregados dois métodos distintos, sendo o primeiro uma análise do comprimento total de vias congestionadas ou com lentidão na cidade a partir da camada de tráfego da plataforma *Google Maps* e o segundo uma análise do volume total de comentários negativos provindos de uma rede social (*Twitter*).

O comprimento total de congestionamento foi obtido a partir do processamento de imagens de mapas de tráfego, coletados a cada 15 minutos na plataforma *Google Maps* durante os dias de 4 a 8 de junho de 2018. Nesses mapas, foi calculado o comprimento das vias de interesse a partir da contagem de pixels das cores vermelho (fluxo lento) e vinho (fluxo congestionado) ao longo de um determinado alinhamento. A conversão da distância na imagem, em pixels, para uma distância real foi feita a partir das transformações características da projeção *Web Mercator*. Em contraste com outros métodos, como a obtenção de informação através da API, o método exposto aqui não possui limitações com relação à quantidade informação obtida. Contudo, cabe ressaltar que nesta abordagem a velocidade média de tráfego só pode ser avaliada de maneira qualitativa a partir das cores dos mapas de tráfego, o que é suficiente para algumas aplicações, como a avaliação da dificuldade geral do motorista em se deslocar no sistema de tráfego, que se relaciona diretamente com a qualidade do sistema.

Para a obtenção da qualidade como percebida pelos usuários, comentários relacionados ao tráfego foram obtidos a partir de contas de usuários do *Twitter* que declaram em seus perfis serem moradores da cidade durante o período de 4 a 8 de junho de 2018. A coleta foi realizada utilizando palavras-chave relacionadas ao tráfego e os comentários foram divididos em três categorias: positivos, negativos e não relacionados ao tráfego, a partir de três tipos diferentes de classificador. Este método se mostrou viável como uma alternativa para a classificação da qualidade do tráfego que não depende da avaliação de parâmetros de tráfego, já que parte diretamente da opinião geral dos usuários.

Chegou-se à conclusão que os dados de comentários negativos relativos ao tráfego seguem uma distribuição horária média com dois picos, um em torno das 9h e outro em torno das 19h, coincidindo com os picos horários observados no carregamento das vias obtidos da plataforma de mapeamento de tráfego. Além disso, foi verificada uma significativa correlação entre a contagem média de *tweets* negativos e o comprimento de vias congestionadas ou com lentidão medidos na plataforma de mapeamento. Foi evidenciada uma dispersão um pouco maior para contagens menores que 10 *tweets*, fato que pode ter sido causado pela natureza discreta dos dados, contudo trabalhos futuros que possam utilizar bases de dados para períodos maiores que o utilizado neste trabalho podem verificar uma redução nessa dispersão já que, em uma análise da média, o ruído contido nos dados tende a ser reduzido.

A obtenção de resultados com distribuições similares a partir de dois métodos diferentes é um forte indicativo de que ambos podem ser empregados para a avaliação da qualidade de tráfego em um sistema urbano, provendo grande volume de dados para a avaliação de fenômenos de tráfego. Trabalhos futuros podem verificar possíveis variações geográficas que possam influenciar a relação entre comentários no *Twitter* e o comprimento de congestionamento obtido da plataforma *Google Maps*, explorando o comportamento em outras grandes cidades.

**Agradecimentos**
Os autores agradecem o apoio da CAPES pelo suporte financeiro, sob a forma de bolsas de mestrado.

**REFERÊNCIAS BIBLIOGRÁFICAS**

Albuquerque F. C. et al (2012). *Extrator de fatos relacionados ao tráfego*. In: SBBD (Short Papers). p. 169-176.

Amin S. et al (2008) Mobile Century Using GPS Mobile Phones as Traffic Sensors: A Field Experiment. 15th World Congress on Intelligent Transportation Systems, p. 8–11, 2008.

Battersby, S. E. et al. (2014) *Implications of web Mercator and its use in online mapping.* Cartographica: The International Journal for Geographic Information and Geovisualization, v. 49, n. 2, p. 85-101, 2014.

Bird S.; Klein E.; Loper E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Collins C.; Hasan S.; Ukkusuri S. V. (2013) *A novel transit rider satisfaction metric: Rider sentiments measured from online social media data.* Journal of Public Transportation, v. 16, n. 2, p. 2.

Gal-Tzur A. et al (2014). *The potential of social media in delivering transport policy goals.* Transport Policy, v. 32, p. 115-123, 2014.

Gong, Y.; Deng, F.; Sinnott, R. O. (2015) *Identification of (near) Real-time Traffic Congestion in the Cities of Australia through Twitter.* In: Proceedings of the ACM First International Workshop on Understanding the City with Urban Informatics. ACM, 2015. p. 7-12.

Goutte C.; Gaussier E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2005. p. 345-359.

Joachims T. (1998). Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. Springer, Berlin, Heidelberg, 1998. p. 137-142.

Luong T. T. B.; Houston, D. (2015). *Public opinions of light rail service in Los Angeles, an analysis using Twitter data.* iConference 2015 Proceedings, 2015.

Mai, E.; Hranac, R. (2013). *Twitter interactions as a data source for transportation incidents.* In: Proc. Transportation Research Board 92nd Ann. Meeting. 2013.

Pan, B. et al. *Crowd sensing of traffic anomalies based on human mobility and social media.* In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2013. p. 344-353.

Pathak, A. et al (2015). *A city traffic dashboard using social network data*. In: Proceedings of the 2nd IKDD Conference on Data Sciences. ACM, 2015. p. 8.

Oliveira T. B. F. *et al.* (2014) *Traffic information extraction from a blogging platform using knowledge-based approaches and bootstrapping.* In: Computational Intelligence in Vehicles and Transportation Systems (CIVTS), 2014 IEEE Symposium on. IEEE. p. 6-13.

Roesslein J. (2009) *tweepy Documentation.* Online] http://tweepy. readthedocs. io/en/v3, v. 5.

Schulz, A.; Ristoski, P.; Paulheim, H. (2013) *I see a car crash: Real-time detection of small scale incidents in microblogs.* In: Extended Semantic Web Conference. Springer, Berlin, Heidelberg, 2013. p. 22-33.

TRB (2010) *Highway Capacity Manual 2010.* Transportation Research Board, National Research Council, Washington, DC, EUA.

Wang, F. e Yanqing, X. (2011). *Estimating O-D Travel Time Matrix by Google Maps API: Implementation, Advantages, and Implications*. Annals of GIS, Vol. 17, n 4, p 199–209, 2011.

Wang, S. et al. (2015) Citywide traffic congestion estimation with social media. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2015. p. 34.