INTER FACES CIENTÍFICAS
EXATAS E TECNOLÓGICAS

ISSN IMPRESSO - 2359-4934
E- ISSN - 2359-4942
DOI - 10.17564/2359-4934.2018v2n3p21-30

GERAÇÃO AUTOMÁTICA DE *TWEETS* SOBRE AS CONDIÇÕES DO TRÂNSITO: UMA ABORDAGEM BASEADA EM *TEMPLATES*

AUTOMATIC GENERATION OF TWEETS ON TRAFFIC CONDITIONS: AN APPROACH BASED ON TEMPLATE

GENERACIÓN AUTOMÁTICA DE TWEETS SOBRE LAS CONDICIONES DE TRÁNSITO: UN ENFOQUE BASADO EN PLANTILLAS

Jonh Paulo Silva Santos[1] Adolfo Pinto Guimarães[2]

RESUMO

Com a popularização da Internet e a facilitação ao acesso à tecnologia, a utilização das redes sociais vem crescendo rapidamente nos últimos anos. No Twitter, por exemplo, são publicadas em média 50 milhões de mensagens por dia, o que o transforma em uma potencial base de dados a ser explorada. Esse trabalho tem por objetivo a geração de *tweets* automáticos sobre a situação do trânsito; o Twitter foi utilizado como uma fonte de informação e através do processo de Mineração de Dados foram extraídas informações sobre as condições do trânsito para a geração desses *tweets*. O resultado desse projeto foi a criação de um programa que publica no Twitter essas mensagens geradas.

PALAVRAS-CHAVES

Mineração de Dados no Twitter. Processamento de Linguagem Natural. Geração de Linguagem Natural. Classificador Bayesiano Ingênuo. ITS.

ABSTRACT

With the popularization of Internet and the easy access to technology, the use of social networks has grown rapidly in recent years. On Twitter, for example, are published an average of 50 million messages a day, which makes a potential database to be explored. This study aims the automatic generation of tweets about the traffic situation; Twitter was used as a source of information and through the data mining process were extracted information on traffic conditions for the generation of these tweets. The result of this project was to create a tool that publishes the generated messages on Twitter.

KEYWORDS

Twitter Data Mining. Natural Language Processing. Natural Language Generation. Naive Bayes Classifier. ITS.

RESUMEN

Con la popularización de Internet y la facilitación al acceso a la tecnología, la utilización de las redes sociales ha crecido rápidamente en los últimos años. En Twitter, por ejemplo, se publican en promedio 50 millones de mensajes al día, lo que lo convierte en una potencial base de datos a explotar. Este trabajo tiene por objetivo la generación de tweets automáticos sobre la situación del tránsito; el Twitter fue utilizado como una fuente de información ya través del proceso de Minería de Datos se extrajo información sobre las condiciones del tránsito para la generación de esos tweets. El resultado de este proyecto fue la creación de un programa que publica en Twitter esos mensajes generados.

PALABRAS CLAVE

Minería de Datos en Twitter. Procesamiento de Lenguaje Natural. Generación de Lenguaje Natural. Clasificador Bayesiano Ingenuo. ITS.

1 INTRODUÇÃO

Com o crescimento da população, o tema transporte urbano tem se tornado um grande desafio para as cidades. Problemas como engarrafamentos, falta de estacionamento, mobilidade urbana, acidentes, qualidade do transporte público têm ganhado destaque nos debates em busca de soluções para essas questões (AN; LEE; SHIN, 2011). Segundo o relatório (ONU, 2012), 80% da população latino-americana vive em centros urbanos e 14% (cerca de 65 milhões) habitam metrópoles como São Paulo e Cidade do México. As cidades da região metropolitana se destacam pela importante participação do transporte público (43%), da caminhada e do ciclismo (28%) nos deslocamentos e no desenvolvimento de sistemas de transporte coletivo integrado (BRT).

O uso de tecnologia em conjunto com telecomunicação e eletrônica, na gestão, fiscalização e operação se mostrou uma alternativa para os problemas citados anteriormente. Conhecidos como Sistemas de Transportes Inteligentes (Intelligent Transport Systems – *ITS*) (BAZZAN; KLÜGL, 2013) tais sistemas têm como objetivo a coleta, compartilhamento, processamento e redistribuição de dados para a melhoria do transporte.

Com uma média de 50 milhões (MEG, 2016) de *tweets* (mensagens do Twitter) publicados por dia sobre diversos temas como política, moda, gastronomia e trânsito, o Twitter se destaca como importante meio de comunicação, colaboração, compartilhamento de ideias e também como uma potencial base de dados a ser explorada. Alguns trabalhos mostram o Twitter como uma importante ferramenta para análise de eventos em tempo real. Por exemplo, Gomide e outros autores (2011) apresenta uma proposta de mapeamento dos casos de dengue baseado nos relatos de textos publicados no Twitter. Já o trabalho proposto por Narayanan, Liu e Choudhary (2009) mostra um sistema que aplica a análise de sentimentos para extrair informações do Twitter a respeito das eleições presidenciais americanas de 2012.

O Twitter também tem se mostrado uma relevante fonte de dados para as aplicações de ITS. Os trabalhos de Ribeiro e outros autores (2012) e Lauand e Oliveira (2013) apresentam diferentes estudos de como os dados do Twitter podem auxiliar a extrair informações do trânsito de duas importantes cidades brasileiras. Já Maghrebi e outros autores (2015) apresentam uma proposta que visa mapear regiões de interesse no trânsito baseado na análise de sentimento extraída a partir dos *tweets*.

Dentro desse contexto, este trabalho propõe a criação de uma ferramenta que analise dados sobre o trânsito, por meio de processamento de linguagem natural e da análise de sentimento dos dados coletados (NARAYANAN; LIU; CHOUDHARY, 2009). A partir disso gera-se uma saída, em forma de texto, sobre a situação do trânsito. Foi desenvolvido um aplicativo que relaciona informações do trânsito extraídas de *tweets* com dados de rotas de trânsito obtidos a partir de geolocalização. São então gerados com o uso de técnicas de geração de linguagem natural baseada em *templates*, *tweets* (mensagens de texto curtas com 140 caracteres) sobre a situação do tráfego em uma determinada localidade. Essas mensagens são publicadas em uma conta do Twitter criada para este propósito.

Os *templates* e demais técnicas de geração de linguagem natural têm sido utilizadas para combinar informações distintas e apresentar um texto coerente ao usuário. O trabalho propõe um sistema que gera descrições de rotas mais próximas à utilização de geração de linguagem natural baseada em *templates* para gerar descrições de rotas mais naturais para o ser humano (TELLES; GUIMARÃES; MACEDO, 2012).

Este trabalho foi dividido da seguinte forma: na Seção 2 é apresentada uma breve introdução sobre os principais conceitos essenciais para o entendimento do trabalho. Na Seção 3 é apresentada a ferramenta desenvolvida: arquitetura, *datasets* e proeminências da metodologia utilizada são detalhadas. Na Seção 4 são exibidos os resultados obtidos com os textos gerados. Por fim, na Seção 5 é mostrada a conclusão do trabalho com uma breve discussão sobre as direções futuras desta proposta.

2 CONCEITOS

A seguir é apresentada uma breve explanação sobre os principais conceitos básicos envolvidos no trabalho. A proposta é fornecer ao leitor um conhecimento básico a respeito dos termos e da fundamentação teórica desenvolvida nas seções seguintes.

2.1 SISTEMAS DE TRANSPORTE INTELIGENTES (ITS)

Segundo Bazzan e Klügl (2013) ITS podem ser visto como um sistema em que as tecnologias de informação e comunicação são aplicadas em áreas relacionadas com a rede de transporte (por exemplo: infraestruturas, os veículos, a gestão do tráfego e mobilidade, e da iteração entre todos esses elementos). O uso de tecnologias destes tipos se torna cada vez mais essencial para o planejamento e gerenciamento dos transportes, uma vez que há um esforço contínuo em tornar mais eficientes os deslocamentos das pessoas – por razões econômicas ou ambientais (PEREIRA, 2007).

Em geral ITS podem ser dividido em cinco grandes áreas: 1 - **ATMS** (Advanced Traffic Management System); 2 - **ATIS** (Advanced Traveller Information System); 3 - **AVCSS** (Advanced Vehicle Control and Safety System); 4 - **CVO** (Commercial Vehicle Operation); 5 - **APTS** (Advanced PublicTransportation System)**.**

Cada uma destas áreas engloba uma série de conceitos e aplicações para resolver os mais diversos problemas relacionados a gerência do trânsito. Do ponto de vista de aplicação, Pereira (2007) apresenta uma classificação de acordo com os objetivos que o projeto fornece:

- **Desenvolvimento de Sistemas de Navegação**: Tem por objetivo fornecer informações aos usuários para que possam planejar os seus deslocamentos – ou até mesmo adaptá-los durante a sua realização – da forma mais eficiente possível;
- **Assistência para a Condução Segura de Veículos**: Consiste na utilização de diversos sensores capazes de fornecer informações tanto dos veículos quanto das vias, possibilitando o envio de alertas conforme o grau de perigo da situação em que os usuários se encontram;
- **Otimização do Gerenciamento de Tráfego**: Esta otimização busca permitir uma atuação mais efetiva dos órgãos gestores no gerenciamento do tráfego, seja para a redução de congestionamentos ou pela capacidade de aplicar medidas para amenizar o impacto de eventos como acidentes e operações especiais de trânsito;
- **Aumento da Eficiência no Gerenciamento de Vias**: Está voltado ao ganho na fluidez dos veículos na rede viária, por meio da identificação das condições das vias e da localização dos veículos.

Este trabalho pode ser classificado dentro da área de **ATMS** já que ao reportar informações sobre a condição do trânsito estamos contribuindo direta ou indiretamente com o gerenciamento do tráfego. Do ponto de vista da aplicação, esse projeto se encaixa dentro do **Desenvolvimento de Sistemas de Navegação** pela razão descrita anteriormente.

2.2 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN) é um campo da inteligência artificial que trabalha com a interação humano-computador do ponto de vista do entendimento da linguagem. Essa área pode ser dividida em duas subáreas distintas, mas, ao mesmo tempo, fortemente relacionadas: (1) a compreensão da linguagem natural e (2) a geração da linguagem natural. Segundo Reiter e Dale (2000), tais conceitos podem ser enxergados como um sendo o inverso do outro. A geração de linguagem natural consiste em mapear uma representação interna do computador em informação na linguagem humana. Já a compreensão de linguagem natural consiste em traduzir a linguagem humana em alguma representação interna para o computador.

Para executar tal processo dentro de um sistema computacional é necessária a combinação de diversas técnicas e algoritmos. Para este trabalho focaremos nos **conceitos de classificação de texto** e **análise de sentimentos** para extrair significados relevantes dos textos.

A geração de linguagem natural pode ser entendida como um tradutor que converte uma representação baseada em computador para uma representação em linguagem natural. Existem diferentes abordagens que vão desde as mais simples – utilizar sentenças predefinidas (cannedtexts) ou modelos (templates) que mesclam textos fixos e variáveis – até as mais complexas – gerar documentos analíticos completos e totalmente dinâmicos (específicos para cada cenário apresentado), utilizando técnicas de inteligência artificial, análise de dados, conhecimentos linguísticos e de domínio (REITER; DALE, 2000).

Uma das vantagens de se trabalhar com abordagens que utilizam textos predefinidos *templates* é a menor necessidade de lidar com as complexidades, envolvendo a gramática do idioma e outras questões relativas às linguagens naturais, como as ambiguidades. Além disso, a utilização de textos pré-definidos faz com que sejam apresentadas sentenças com uma forma de escrita mais natural.

2.3 CLASSIFICAÇÃO DE TEXTO E ANÁLISE DE SENTIMENTO

A metodologia utilizada para o desenvolvimento da ferramenta tomou como base os conceitos de descoberta do conhecimento já consolidados na literatura. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o método tradicional de se transformar conhecimento em dados baseia-se na análise manual e interpretação dos dados. Neste contexto surge o KDD (Knowledge Discovery in Databases), que é segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o processo de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados.

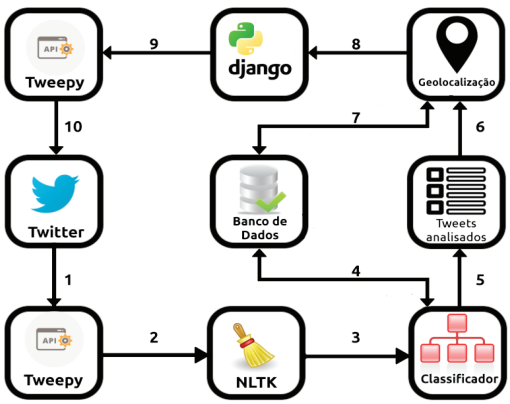
A classificação é uma das tarefas populares dentre do KDD (GOLDSCHMIDT; PASSOS, 2005). O processo de classificação refere-se à tarefa de associar classes (ou categorias) a um determinado conjunto de dados. A partir desta base classificada é possível construir um modelo capaz de categorizar automaticamente instâncias não conhecidas. Existem diversas técnicas de classificação na literatura e dentre as utilizadas pode-se citar: o algoritmo C4.5, as Redes Neurais MLP (multilayerperceptron), o algoritmo K-NN (k-nearestneighbors), os classificadores Bayesianos e algoritmos genéticos.

Uma aplicação da classificação é denominada de análise de sentimento. A análise de sentimento pode ser definida como uma maneira de analisar a opinião e os aspectos sociais, psicológicos, filosóficos e comportamentais de uma pessoa ou grupo em uma situação específica sobre produtos, serviços, eventos, pensamentos, pessoas ou organizações em uma situação particular. Sendo assim, analisar sentimentos implica em identificar como os sentimentos são expressados em textos e se essas expressões revelam opiniões positivas ou negativas sobre um determinado sujeito ou objeto (NASUKAWA et al., 2003).

3 GERAÇÃO AUTOMÁTICA DE TWEETS SOBRE AS CONDIÇÕES DO TRÂNSITO

Para atingir os objetivos descritos anteriormente foi desenvolvida uma ferramenta capaz de coletar informações do Twitter, analisar os textos coletados e gerar a partir desta análise novos textos de 140 caracteres (tweets) que representem o estado do trânsito e para isso foi criado um perfil no Twitter que publica de forma automática as informações produzidas. Para que o usuário possa obter as informações geradas pelo sistema, ele precisa acessar o perfil **@TransitoAJU_** no Twitter e verificar as mensagens publicadas nesse perfil. A Figura 1 mostra a arquitetura da ferramenta desenvolvida.

Figura 1 – Arquitetura da Aplicação



Fonte: Elaborada pelos autores.

Primeiro, os *tweets* foram obtidos utilizando a API *Tweepy* para a análise das condições do trânsito no espaço de tempo de uma hora. Após isso eles foram pré-processados pelo NLTK, são eliminadas, por exemplo, as palavras que não são importantes na análise sobre o sentimento que o *tweet* transmite sobre o trânsito (artigos, preposições, URL etc.). Uma vez que os *tweets* foram pré-processados, eles são classificados por um classificador (nessa abordagem utilizamos um classificador Bayesiano ingênuo) em **positivo**, **negativo** e **neutro** de acordo com o sentimento expresso neles. Para essa tarefa, o classificador consultava a base de dados que contém uma série de *tweets* já classificados, comparando as informações extraídas dos novos *tweets*, com as informações da base.

Após esse passo, as informações extraídas são associadas aos dados de trânsito obtidos a partir de geolocalização do mesmo espaço de tempo dos *tweets* (uma hora), dados estes que também estão na base de dados. A partir disso é feita uma análise dos dados e gerada uma mensagem sobre as condições do trânsito a partir de um *template* que melhor se adéqua a mensagem (esse template será preenchido pelo framework Django com as informações da mensagem). Por fim, a mensagem gerada é enviada ao Tweepy para a publicação na *timeline* do perfil **@TransitoAJU_**.

3.1 DATASET

O trabalho é composto por dois *datasets*: (1) base de textos coletados do Twitter e (2) base que representa dados de geolocalização. Os textos foram coletados a partir de perfis oficiais dos órgãos de transporte que publicam sobre as condições de trânsito locais e de usuários que citavam tais perfis para reportar algum problema. A base possui **1009** *tweets* que foram classificados manualmente em negativos (que reportam problemas. Ex.: Colisão na rua Y), positivos (que reportam situação de fluidez. Ex.: Trânsito tranquilo na Rua Z) e neutros (tweets que não continham informações sobre a situação do trânsito. Exemplo: Motorista esperto faz sempre revisão). No total, **510** foram categorizados como negativos, **227** como positivos e **272** como neutros.

O outro *dataset* consiste em dados de geolocalização. O grande problema em se trabalhar com dados somente do Twitter é que dificilmente os usuários informam quando o trânsito está bom. Normalmente, uma mensagem no Twitter é publicada quando existe algum tipo de problema e para contornar a falta de informações positivas sobre o trânsito foi proposta uma abordagem que combinasse os textos com dados de geolocalização capazes de informar a fluidez do trânsito nas áreas analisadas.

No entanto, a obtenção desta base foi um dos principais obstáculos enfrentados durante esse projeto. Tais dados poderiam ser obtidos por meio de órgãos do governo que medem a velocidade das vias com o auxílio de sensores, por exemplo. A primeira ideia foi a utilização dos dados de uma fonte que os captasse em tempo real, mas isso não foi possível uma vez que o órgão responsável da cidade estudada (Aracaju-SE) não disponibiliza tais informações.

Para a validação da implementação da ferramenta, os dados de geolocalização foram inseridos manualmente na base da seguinte forma: 1 - Dentre os *tweets* da base de textos coletados, foram identificados os endereços mais citados; 2 - Com a API do *Google Maps* é possível exibir a situação do trânsito num determinado ponto, então passando-se as coordenadas dos endereços mais citados, foi possível obter um mapa com a situação do trânsito naquele local; 3 - O passo seguinte foi analisar as imagens para inserir no banco as informações relativas à situação do trânsito.

A Figura 2 mostra exemplos dos dados coletados a partir das imagens. A coluna **sit** representa a situação do trânsito no instante em que o dado foi coletado: **i** – intenso, **m** – moderado, **l** – leve.

Figura 1 – Exemplos de dados de geolocalização

latitude	longitude	data_hora	local	sit
-10.951.196	-37.051.837	10/11/2015 18:00	Av. A	i
-10.951.196	-37.051.837	10/11/2015 18:10	Av. A	i
-10.922.700	-37.057.273	10/11/2015 20:00	Av. B	m
-10.922.700	-37.057.273	10/11/2015 20:10	Av. B	m
-10.949.294	-37.070.721	10/11/2015 21:30	Av. C	l
-10.949.294	-37.070.721	10/11/2015 21:40	Av. C	l

Fonte: Dados da pesquisa

Foram coletados **1872** registros de geolocalização: **647**mostravam trânsito leve, **545** mostravam trânsito moderado e **680** mostravam trânsito intenso, totalizando **1872** registros na base de dados. Dessa forma foi formada a base com dados de geolocalização sobre o trânsito para o projeto. É válido observar que como esses dados não foram obtidos em tempo real, os *tweets* do Módulo 2 foram gerados de forma retroativa, de acordo com essa base. Isso foi feito para validar a proposta da ferramenta, porém a ausência desses dados em tempo real é a principal limitação deste trabalho e será tratada em trabalhos futuros.

3.2 METODOLOGIA

A seguir são apresentados os passos para atingir o objetivo proposto. A ferramenta foi implementada em Python, utilizando diversas API para a implementação e integração dos componentes descritos na arquitetura da Figura 1. O desenvolvimento da ferramenta pode ser dividido em dois módulos referenciados como **Módulo A** e **Módulo B**. O primeiro módulo considera apenas os dados extraídos do Twitter. Já o segundo, utiliza, além desses dados, as informações de geolocalização extraídas do *Google Maps*. Antes de tais etapas fez-se necessário a implementação do classificador de textos.

O primeiro passo foi treinar um classificador com os dados obtidos do Twitter. Para a tarefa de classificação foi utilizado o *Natural Language Toolkit* (NLTK) que é uma plataforma para a construção de programas em Python que trabalham com dados em linguagem natural. Foi utilizado o classificador Bayesiano ingênuo que funciona bem para atividades deste trabalho, no entanto, a implementação e comparação de outros classificadores devem ser levados em consideração e é proposto como direção futura deste trabalho.

O classificador foi testado a partir de 9 *tweets* inseridos manualmente. Como 9 *tweets* não apresentam um conjunto de dados relevantes para uma tarefa de classificação, foram coletados 1000 outros *tweets*. Essa tarefa foi dividida em duas etapas: (1) utilizar o classificador criado para classificar de forma automática mil novos *tweets*, aumentando assim a base de dados; (2) revisar manualmente esses novos **um mil** *tweets* para verificar a eficiência do classificador construído anteriormente. Com isso, a **base de treinamento** do classificador é composta por **1009** (um mil e nove) *tweets*: 1000 do segundo passo e 9 do primeiro passo sendo, **510** negativos, **227** positivos e **272** neutros.

Após a coleta dos dados, foi preciso pré-processar esses *tweets* para extrair apenas informações importantes. Essa etapa envolve:

- Converter os *tweets* para letras minúsculas;
- Como a intenção não é acessar as URL que possam existir nos *tweets*, identifica-las por meio de expressão regular e eliminá-las;
- Também por meio do uso de expressão regular, reconhecer e eliminar os nomes de usuários (@ username);
- *Hashtags* podem transmitir alguma informação útil, por isso é preciso mantê-las, substituindo pela mesma palavra sem o **#**. Por exemplo, **#acidente** substituído por **acidente**;
- Remover pontuação no início e no final dos *tweets*. Por exemplo: "o trânsito está bom!" pode ser substituído por "o trânsito está bom". Também é necessário substituir vários espaços em branco com um único espaço em branco;
- *Stopwords* são palavras que não indicam qualquer sentimento e podem ser removidas (artigos e preposições, por exemplo);
- Letras repetidas – nos *tweets*, às vezes as pessoas repetem as letras para salientar a emoção. Por exemplo. "leeeento", "leeeeeeeeeento" podem ser substituídas por "lento". Considerar que duas ou mais letras repetitivas em palavras serão substituídas por uma apenas;
- As palavras devem começar com uma letra, para simplificar o processamento. Desta forma, pode-se remover todas aquelas palavras que não começam com uma letra. Por exemplo, remover horas nos *tweets* "19h34m".

O resultado desse pré-processamento é um vetor de características utilizado na implementação e treinamento do classificador. O vetor de características é usado para construir um modelo em que o classificador aprende com os dados de treinamento e posteriormente pode ser usado para classificar os dados inéditos.

Com o classificador pronto e a base aumentada e revisada foi criado o **Módulo A**. Esse módulo consiste em uma ferramenta que de forma automática, classifica um conjunto de *tweets* e gera uma mensagem sobre as condições do trânsito a partir do conjunto analisado. Para gerar esse texto foram utilizadas técnicas de Geração de Linguagem Natural a partir de *templates*. Foram criados dois modelos de *templates* nessa etapa:

- Trânsito fluindo tranquilo. {{ qtdade }} *tweet*(s) reportando;
- Trânsito ruim. {{ qtdade_lento }} report(s) de trânsito lento, {{ qtdade_atrope }} report(s) de atropelamento, {{ qtdade_batidas }} reports de batida.

Dentro dos {{ }} são inseridas informações sobre o trânsito obtidas a partir da análise do texto. No período de uma hora, um novo conjunto de *tweets* é coletado. Esses textos são enviados para o classificador que os classifica em negativos, positivos ou neutros. A depender da classificação da maioria dos *tweets*, um deste *template* é selecionado, preenchido e publicado no perfil **@TransitoAJU_**. Com a análise dos resultados desta etapa, percebeu-se que poucos *tweets* positivos eram reportados pelos usuários. Desta forma, foi desenvolvido o Módulo B do projeto que agrega informações de geolocalização para a escolha do *template*.

Em média, 70% dos usuários que mencionavam os perfis escolhidos reportavam apenas acidentes, colisões, trânsito lento, ou seja, *tweets* ruins sobre a situação do trânsito. Para resolver esse problema, foram utilizados dados obtidos a partir de geolocalização sobre a situação do trânsito e a partir desse ponto, no processo da geração dos *tweets*, os dados de geolocalização também foram analisados, aumentando a fonte de informação sobre o trânsito.

Nessa etapa do projeto foi incluída a análise dos dados de geolocalização no processo de geração dos *tweets*, uma vez que com o *Tweepy* é possível recuperar vários metadados sobre o *tweet*: data, hora, perfil, id etc. Dentre esses metadados, é possível obter também a latitude e longitude da onde o *tweet* foi originado. Com a latitude e longitude, é possível cruzar essas informações com a base de dados de geolocalização na hora da análise dos dados para geração do *tweet* sobre o trânsito.

Como a quantidade da informação para geração aumentou, o número de *templates* também aumentou. Foram adicionados os seguintes modelos:

- Trânsito ruim. Acidente na {{ localizacao }};
- Trânsito ruim. Acidente na {{ localizacao }};
- Fique atento, queda de árvore na {{ localizacao }};

- Trânsito lento na {{ localizacao }};
- Trânsito na {{ localizacao }} fluindo bem;
- ALERTA. Acidente com vítimas na {{ localizacao }};
- Trânsito tranquilo na {{ localizacao }}

O funcionamento deste módulo assemelha-se ao anterior. No entanto, a escolha do *template* leva em consideração a análise de sentimento a partir do classificador, as palavras extraídas do texto e os dados de fluidez do trânsito a partir da base de geolocalização.

4 RESULTADOS

No primeiro módulo foram analisados **437** *tweets*, classificados da seguinte maneira: **233** negativos, **122** positivos e **82** neutros, constituindo assim a primeira **base de testes**. Na segunda etapa com a utilização dos dados de geolocalização foram analisados **671** *tweets*, e classificados como segue: **301** negativos, **263** positivos e **107** neutros, constituindo assim a segunda **base de testes**.

Foi possível observar que devido a limitação de informações e consequentemente as poucas opções de *templates* utilizados, os textos gerados pelo **Módulo A** são mais simples, informando apenas se o trânsito está bom ou ruim. A quantidade de eventos é extraída a partir das palavras que mais frequentemente apareceram nos textos.

Já a segunda abordagem (**Módulo B**) permitiu a geração de *tweets* mais elaboradas, uma vez que os dados de geolocalização passaram a ser levados em consideração durante a análise da situação do trânsito e com o melhoramento dos *templates* utilizados, os resultados foram *tweets* mais elaborados, com um grau de informação maior.

5 CONCLUSÕES E TRABALHOS FUTUROS

Os primeiros resultados foram obtidos sem a aplicação dos dados de geolocalização para a determinação da situação do trânsito e podem ser analisados da seguinte forma: do ponto de vista do **processo**, as mensagens geradas nessa etapa foram satisfatórias visto que nesse modelo não houve problemas para se obter/tratar os *tweets* e gerar a mensagem final para o usuário. Já do ponto de vista da **qualidade da mensagem gerada**, essa etapa pode ser dita como imprecisa, uma vez que o *tweets* gerados tinham informações mais limitadas, por exemplo, os *tweets* gerados não tinham as localizações onde os fatos (acidentes, congestionamentos) estavam ocorrendo.

Os resultados obtidos com a aplicação dos dados de geolocalização podem ser analisados de forma inversa: do ponto de vista do **processo**, houve problemas para a obtenção dos dados de geolocalização em tempo real, o que tornou a geração da mensagem mais trabalhosa quando comparada ao passo anterior. Já em relação a **qualidade da mensagem gerada**, essa etapa pode ser dita como mais precisa que a anterior visto que a quantidade de informações para gerar os *tweets* era maior e os modelos de *templates* mais diversificados.

Uma proposta para melhoria deste trabalho é a automatização de extração de dados de geolocalização. Alternativas devem ser analisadas, dada a dificuldade em se obter estas informações em tempo real dos órgãos públicos de gerência do trânsito. Outra melhoria é obtenção de mais dados. Atualmente a base conta com 1009 textos de trânsitos coletados. Isso pode ser feito coletando mais dados do Twitter e com a integração de outras fontes de dados. A técnica de geração de texto utilizada é relativamente simples, o que acarreta em um texto inflexível. Outras técnicas de Geração de Linguagem Natural podem ser estudadas daqui em diante.

REFERÊNCIAS

AN, S.; LEE, B.H.; SHIN, D.R. A Survey of Intelligent Transportation Systems. **2011 Third International Conference on Computational Intelligence, Communication Systems and Networks**, p. 332–337, 2011.

BAZZAN, A. L. C.; KLÜGL, F. Introduction to Intelligent Systems in Traffic and Transportation. **Synthesis Lectures on Artificial Intelligence and Machine Learning**, v. 7, n. 3, p. 1–137, 2013.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v.17, n.3, p.37, 1996.

GOLDSCHMIDT, R.; PASSOS, E.L. **Data mining. Um guia prático**. Elsevier ed.
GOMIDE, J. et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. **Proceedingsofthe ACM WebSci'11,** Koblenz: Germany, june 14-17, 2011, p.1-8.

LAUAND, B.P.; OLIVEIRA, J. TweeTraffic: ferramenta de análise das condições de trânsito baseado nas informações do Twitter. **II Brazilian Workshop on Social Network Analysis and Mining**, 2013, p.1-6.

MAGHREBI, M. *et al.* Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time. **IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC**, octob. 2015. p.208-213.

NARAYANAN, R.; LIU, B.; CHOUDHARY, A. Sentiment analysis of conditional sentences. **Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 EMNLP 09**, n. August, p.180, 2009.

NASUKAWA, T. *et al*. Sentiment analysis: Capturing favorability using natural language processing. **Proceedings of the 2nd international conference on Knowledge capture**, 2003. p.70-77.

ONU. **Estado de lasCiudades de América Latina y el Caribe 2012, Rumbo a una nuevatransición urbana**.

PEREIRA, W.F. **O uso de sistemas inteligentes para o aumento da eficacia do transporte público por ônibus**. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2007.

REITER, E.; DALE, R. **Building Natural Language Generation Systems**. Cambridge University Press, 2000.

RIBEIRO, S.S. *et al*. Traffic observatory: a system to detect and locate traffic events and conditions using Twitter. **Proceedings of the 5th International Workshop on Location-Based Social Networks - LBSN '12**, 2012. p.5.

TELLES, R.; GUIMARÃES, A.; MACEDO, H. Automated feeding of POI base for the generation of route descriptions. **Proceedings of the 6th Euro American Conference on Telematics and Information Systems - EATIS '12**, 2012. p.253.

Recebido em: 10 de Março de 2017
Avaliado em: 22 de Março de 2017
Aceito em: 21 de Abril de 2017

1 Graduado em Ciência da Computação pela Universidade Tiradentes – UNIT. E-mail: jonhpaulo.jp@gmail.com

2 Mestre em Ciência da Computação pela UFMG; Professor Adjunto da Universidade Tiradentes – UNIT.E-mail: adolfoguimaraes@gmail.com