UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CENTRO DE INFORMÁTICA

2012.2

UBIBUSANALYSIS – UMA FERRAMENTA DE INTERPRETAÇÃO DE MENSAGENS DE TRÂNSITO COM ANÁLISE DE SENTIMENTOS

VANESSA GOMES DE LIMA

TRABALHO DE GRADUAÇÃO

RECIFE,

30 DE ABRIL DE 2013

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

VANESSA GOMES DE LIMA

UBIBUSANALYSIS – UMA FERRAMENTA DE INTERPRETAÇÃO DE MENSAGENS DE TRÂNSITO COM ANÁLISE DE SENTIMENTOS

TRABALHO DE CONCLUSÃO DE CURSO APRESENTADO PELA ALUNA VANESSA GOMES DE LIMA, SOB A ORIENTAÇÃO DA PROFESSORA ANA CAROLINA SALGADO, À COORDENAÇÃO DE CIÊNCIA DA COMPUTAÇÃO DA UNIVERSIDADE FEDERAL DE PERNAMBUCO, COMO REQUISITO PARA OBTENÇÃO DO TÍTULO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO.

RECIFE,

30 DE ABRIL DE 2013.

DEDICO ESTE TRABALHO AOS MEUS FAMILIARES, EM ESPECIAL AOS MEUS PAIS LUCINEIDE ANA DE LIMA GOMES E LEÔNIDAS DA SILVA GOMES.

AGRADECIMENTOS

Agradeço primeiramente à Deus, por me dar saúde para estudar, coragem para aguentar e sabedoria para continuar. Consegui ingressar numa instituição pública e isso com certeza mudou minha vida.

Agradeço aos meus pais, por me apoiarem em todas as decisões que tomei até hoje e ficarem felizes a cada vitória que conquisto. A todos os meus familiares, que estão sempre na torcida e vibrando com os meus avanços. Ao meu querido Mateus, que tem sido sempre um grande companheiro.

À toda a equipe do Ubibus, especialmente ao pessoal da Iniciação Científica que sempre me ajudou quando precisei ao longo deste trabalho. À minha orientadora Ana Carolina Salgado – querida Carol -, que foi uma verdadeira mãe na faculdade nos últimos dois anos. Obrigada por aconselhar, apoiar, guiar e ensinar.

Aos meus queridos velhos amigos, que mesmo na ausência estiveram torcendo por mim.

E por fim, e não menos importante, aos amigos que fiz ao longo do curso, fora ou dentro da faculdade. Sei que nossa amizade vai muito além: Débora, Isa, Renato, Dragão, Gabriel, Paula, Hélio, Brhenna, todo mundo do quéisso, galera do CIn, valeu, pessoal!

RESUMO

É possível perceber que o cenário atual do trânsito nas grandes cidades vem piorando a cada dia que passa. Observando a realidade brasileira, os congestionamentos tornam-se cada vez mais frequentes e isso ocorre devido ao grande número de veículos nas vias, acidentes, alagamentos ou outras situações não previsíveis. Propostas de solução fazem parte do projeto Ubibus, que tem como objetivo o desenvolvimento de um sistema de transporte público inteligente, ubíquo e sensível ao contexto. O Ubibus coleta informações de trânsito de diversas fontes, com o objetivo de saber situações das vias e de agregar informação ao usuário final. Porém, não é suficiente apenas capturar tais informações, é preciso que seja feita uma abordagem que possa extrair informação dos textos. Este trabalho de graduação, denominado UbibusAnalysis, tem como objetivo a criação de um sistema Web de captura e análise de informações contextuais extraídas de redes sociais. A análise parte do princípio de coletar mensagens e criar ocorrências de trânsito a partir delas, extraindo positividade ou negatividade de acordo com a situação real das vias relatada nas redes sociais.

Palavras chave: Informações Contextuais, Análise de Sentimentos, Análise de Trânsito, Sistemas Inteligentes de Transporte.

ABSTRACT

The current traffic scenario in large Brazilian cities is getting worse every day. Observing Brazilian's reality, congestion become increasingly frequent and this is due to the large number of vehicles on the streets, accidents, water logging or other unpredictable situations. Some solutions are proposed by the Ubibus Project: an Intelligent Transportation System, ubiquitous and context-sensitive. The Ubibus collects traffic information from various sources, aiming to learn airway situations and aggregate information to the end user. But it is not enough on that system just to capture such information, an approach is needed to extract information from texts. This work, called UbibusAnalysis, aims to create a web capture and analysis of contextual information extracted from social networks. UbibusAnalysis collect messages and create traffic events from them, extracting positive or negative occurrences according to the real situation of the roads.

Keywords: Contextual Information, Sentiment Analysis, Traffic Analysis, ITS.

Índice de Figuras

Sumário

1. Introdução

Neste capítulo são apresentados o contexto e a motivação do trabalho, os principais objetivos além da estrutura do documento.

1.1 Contexto e Motivação

É possível perceber que o cenário atual do trânsito nas grandes cidades vem piorando a cada dia que passa [Zhang, 2011]. Observando a realidade brasileira, isso acontece especialmente onde há um aumento do número de carros particulares nas vias. Os congestionamentos tornam-se cada vez mais frequentes e isso ocorre devido ao grande número de veículos nas vias, acidentes, alagamentos ou outras situações não previsíveis.

Com a Copa do Mundo de Futebol em 2014 e as Olimpíadas em 2016, o Brasil terá que solucionar este problema. O Ubibus, um sistema de transporte público inteligente, ubíquo e sensível ao contexto [VIEIRA et al. 2011], tem como objetivo o desenvolvimento de soluções para esse problema.

O Ubibus coleta informações e mensagens de trânsito de diversas fontes, com o objetivo de saber situações das vias e de agregar informação ao usuário final. Existem, disponíveis na Web, informações contextuais, que são o tipo de informação que se referem ao contexto de uma entidade em relação à localização, tempo, entre outros [ZIMMERMANN et al. 2007].

No conjunto de informações que o Ubibus utiliza, encontram-se as informações contextuais, utilizadas devido a sua facilidade de captura na Web, por serem abundantes e crescentes, por serem postas em tempo real e de qualquer lugar, por serem redundantes e relevantes, entre outros [Magalhães, 2008].

Porém, não é suficiente apenas capturar tais informações, é preciso que seja feita uma abordagem que possa extrair mais **informação** dos textos, a fim de responder perguntas do tipo "Como está a Avenida Caxangá às 8h?", que é uma informação possível de ser obtida a partir de blogs, *microblogs* e redes sociais [Pak, 2010]. Há, portanto, um espaço aberto para contribuição com a análise de mensagens de trânsito provenientes de redes sociais para influência na sugestão de rotas de ônibus.

Temos, então, a Análise de Sentimentos (AS) - ou mineração de opinião - que é uma área recente da Computação que estuda opiniões, sentimento, avaliações e emoções que possam ser expressas em forma de texto [Liu 2010]. Com isso em mente, a abordagem com técnicas de Análise de Sentimentos parece se adequar ao tipo de questionamento levantado, sobre a análise de mensagens de trânsito.

1.2 Objetivos

Este trabalho de graduação, denominado UbibusAnalysis, tem como objetivo a criação de um sistema Web de captura e análise de informações contextuais extraídas de redes sociais. A abordagem utilizada para interpretação das mensagens será por técnicas de Análise de Sentimentos, visto que mensagens de trânsito tendem a ter cunhos positivos ou negativos bem definidos.

O UbibusAnalysis funcionará como um componente do sistema Ubibus [VIEIRA et al. 2011] e o resultado da análise feita por esse novo componente irá povoar a base de dados do Ubibus.

Serão coletados dados sobre Endereços do Grande Recife, pois esses vão ser os dados utilizados pelo UbibusAnalysis. Será, também, construído um Extrator de Informações Contextuais, responsável por buscar nas redes sociais informações sobre ocorrências de trânsito no Recife.

Após essas etapas, o trabalho consiste em desenvolver um módulo de análise de sentimentos, responsável por analisar as mensagens das redes sociais e extrair ocorrências e localidades dos textos, com o objetivo de popular a base de dados do Ubibus [Vieira et al. 2011].

Por fim, será desenvolvida uma API REST capaz de prover serviços Web a serem disponibilizados a diversas aplicações. Esses serviços serão relativos às localidades e ocorrências extraídas das mensagens das redes sociais.

1.3 Estrutura

Além deste capítulo introdutório, com contextualização, motivação e objetivos, o trabalho tem a seguinte estrutura:

No capítulo 2, é feita uma fundamentação conceitual. São apresentados conceitos relacionados ao trabalho e importantes para o entendimento completo da ferramenta. São introduzidos conceitos sobre Contexto Computacional, Sistemas Inteligentes de Transporte, Redes Sociais e Análise de Sentimentos.

No capítulo 3, são detalhados dois estudos relacionados, além de outros trabalhos que usam técnicas de análise de sentimentos para interpretação de mensagens provenientes de *microblogs*.

Os capítulos 4 e 5, respectivamente, apresentam o UbibusAnalysis em termos de arquitetura e implementação, além de uma avaliação dos resultados obtidos. Ao final temos um capítulo de conclusão que também traz perspectivas para trabalhos futuros.

2. Fundamentação Conceitual

No presente capítulo serão fundamentados tópicos pertinentes ao entendimento do trabalho como um todo. Será mostrada uma apresentação sobre Sistemas Inteligentes de Transporte e sobre Redes Sociais.

Após a fundamentação a ser apresentada nessas duas seções, será possível perceber que existe uma demanda por informação de Redes Sociais, mais especificamente no contexto do tráfego urbano, por meio dos Sistemas Inteligentes de Transporte.

Com essa produção e demanda, surge então a necessidade de uma forma de interpretar e extrair informação desses textos oriundos de Redes Sociais. Será mostrada então uma visão sobre Análise de Sentimentos, como o trânsito pode estar relacionado com a Análise de Sentimentos, alguns conceitos básicos para o entendimento do problema, as etapas que formam a análise de sentimentos e os desafios e limitações enfrentadas nessa área.

2.1. Contexto Computacional

Contexto computacional pode ser definido como um conjunto de condições e influências relevantes à aplicação e que tornam uma situação única e compreensível [BRÉZILLON 1999]. De acordo com [ZIMMERMANN et al. 2007], qualquer informação que descreva o contexto de determinada entidade faz parte de uma das seguintes categorias: localização, individualidade, tempo e relações.

Com o uso do contexto computacional diversos aplicativos inteligentes podem ser criados, capazes de se adaptar a determinada situação conforme o conjunto de informações contextuais válidas e suas interações [BALDAUF 2007].

Nas seções a seguir, apresentamos o conceito de Sistemas Inteligentes de Transporte, e logo após um APTS ubíquo e sensível ao contexto, denominado Ubibus.

2.2. Sistemas Inteligentes de Transporte

Nesta seção, fala-se sobre os Sistemas Inteligentes de Transporte (SIT), um exemplo de SIT e sobre o tipo de informação pode ser usada para o desenvolvimento desse tipo de sistema.

2.2.1. O que são SIT?

Os SIT são sistemas que utilizam tecnologias de processamento de informação e comunicação, sensoriamento, tecnologia de controle e navegação aplicados à melhoria do gerenciamento e operação dos sistemas de transporte em geral [AN et al. 2011]. Os SIT ainda se preocupam com a melhoria da segurança viária, o aumento da mobilidade, a redução de custos sociais e até impactos ambientais [SILVA 2000].

De acordo também com o trabalho de [GÓMEZ et al. 2009], Sistemas Inteligentes de Transporte têm por objetivo aplicar tecnologia e melhorar a qualidade dos sistemas de transporte. Uma das subáreas dos SIT são os chamados Sistemas de Transporte Público Avançado (APTS, do inglês *Advanced Public Transportation Systems*), que são voltados ao transporte público. Nesta categoria se inserem aplicações que consistem em prover informações aos passageiros como, por exemplo, tempo de espera na parada e rotas de ônibus. Estas

informações ajudam os usuários a definir seus trajetos e planejar melhor os deslocamentos [SUSSMAN 2005].

2.2.2. Ubibus

O Ubibus tem o objetivo de facilitar o dia a dia das pessoas que utilizam transporte público, oferecendo acesso inteligente a informações de transporte público aos passageiros, em tempo real, baseado em informações dinâmicas de contexto relacionadas aos próprios meios de transporte [VIEIRA et al. 2011]. A arquitetura geral do Ubibus é mostrada na Figura 1, no entanto será chamada a atenção apenas para a *camada de aquisição.*

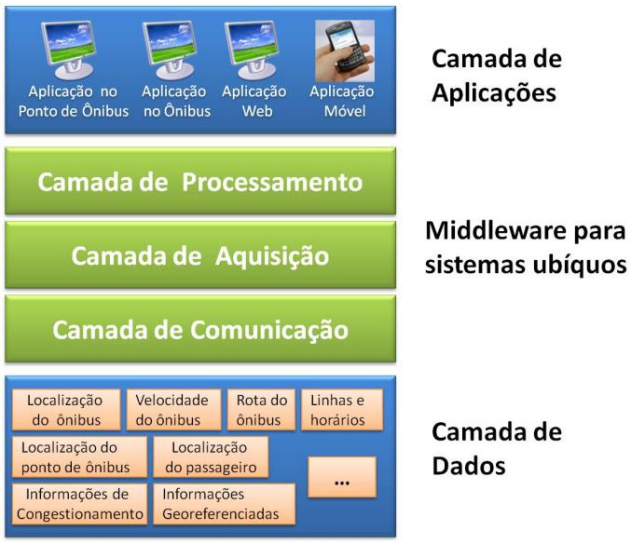


Figura 1 - Arquitetura do Ubibus

É descrito no trabalho de [VIEIRA et al. 2011], que a *camada de aquisição* do Ubibus é a responsável por reunir informações contextuais de diferentes fontes, encaminhando-as para a *Camada de Dados*. No Ubibus, as informações contextuais são adquiridas de fontes como redes sociais (e.g *Twitter*), GPS, câmeras de monitoramento, entre outras. Tais informações podem ser dinâmicas (e.g localização dos ônibus) ou inferidas (e.g presença e intensidade de congestionamento). Os usuários podem ainda utilizar o sistema W*eb* ou dispositivo móvel para adicionar informações sobre seu contexto atual ou de contextos anteriores pelo qual tenham estado.

Existe também uma aplicação do Ubibus, chamada UbibusRoute [LIMA et al. 2012], que é responsável pela comunicação com usuários de transporte coletivo por ônibus, via dispositivo móvel, e que usa informações provenientes de redes sociais para recomendar rotas a esses usuários apoiando-os em suas tomadas de decisão.

2.3. Redes Sociais

Redes Sociais são “serviços baseados na Web que permitem a indivíduos construir um perfil público ou *semi-público* dentro de um sistema limitado, articular uma lista de outros usuários com quem eles compartilham uma conexão, e ver e percorrer suas listas de conexões e aquelas feitas por outras pessoas dentro do sistema” [BOYD & ELLISON 2011]. Esta citação mostra o destaque que a interação entre os usuários exerce no ambiente das redes sociais.

Existem várias Redes Sociais espalhadas pela Internet e o objetivo de cada uma delas pode variar, surgindo algumas diferenças em relação ao público-alvo. A seguir, dois tópicos que falam sobre a postura do usuário na *Web atual* e o *Twitter*, rede de sucesso que tem atraído interesse de usuários e organizações.

2.3.1. A nova postura do usuário na Web

A era das Redes Sociais trouxe, como principal consequência, a mudança de comportamento do usuário perante a rede. Com o paradigma inovador da *Web* 2.0, novas formas de interação surgiram e permitiram aos usuários experiências interessantes.

Então, com a chegada das Redes Sociais, o usuário deixou de ser somente um espectador da Web. Do *status* de *consumidor de informação*, ele passou à condição de *produtor direto de informação*.

Nas Redes Sociais, especificamente, os usuários têm grande liberdade para gerar novo conteúdo, e interagir com outras pessoas. A possibilidade de trocar mensagens entre outros membros da rede, adicionar novos amigos e integrar-se a comunidades virtuais foram exemplos de novas experiências vivenciadas pelo usuário [SILVA FILHO 2011].

As novas opções de integração de usuários geraram o conceito chamado *colaboratividade*. O termo diz respeito à colaboração de diferentes usuários, possivelmente em locais físicos distantes, em prol da elaboração de um conteúdo qualquer.

2.3.2. Twitter

A era atual é a da informação instantânea, onde notícias são produzidas de forma cada vez mais rápida e em maior quantidade. Devido a essa demanda, as notícias precisam também ser transmitidas de maneira rápida, pois os usuários estão cada vez mais acessando mais informações e estão exigindo que estas estejam de forma *enxuta*.

Com toda a popularidade então da Web colaborativa e das redes sociais, surgiu o *Twitter* em 2006. O *Twitter* é uma rede social e *microblog* que permite aos seus usuários postagens em tempo real, chamados *tweets.*

Os denominados *tweets* são mensagens curtas e restritas a um tamanho de 140 caracteres [AGARWAL et. Al. 2011]. *Retweets* são mensagens onde um usuário repete o que outro falou. *Mentions* são *tweets* do tipo resposta de um usuário para outro.

A publicação de informações no *Twitter*, embora que em mensagens curtas, acontece de forma muito rápido. Talvez este seja um dos pontos que expliquem seu sucesso.

Um dos pontos que tem feito o *Twitter* adquirir muito sucesso é o modo de publicação de suas informações. As mensagens são curtas e o modo de publicação é instantâneo. Neste ambiente, os fatos e notícias estão disponíveis em tempo real [RUFINO, 2009]. Mesmo com a limitação no tamanho da mensagem, é muito comum a inserção de links. E as mensagens vão além, proporcionando informações de trânsito, tempo e situações adversas de uma cidade, por exemplo.

2.4. Análise de Sentimentos

Análise de Sentimentos (AS) - ou mineração de opinião - é uma área recente da Computação que estuda opiniões, sentimentos, avaliações e emoções que possam ser expressas em forma de texto [LIU 2010.a]. Dessa forma, pode-se definir o objetivo da AS como sendo identificar o sentimento que usuários apresentem sobre produtos, empresas e até outras pessoas, baseado em conteúdo na Internet [PANG & LEE 2008].

Nos últimos anos, a área de AS tem atraído muita atenção não somente por parte da academia, mas também por parte do mercado. A análise do conteúdo da Web tem se tornado importante devido ao aumento na comunicação via Internet como email, sites e fóruns [ABBASI et al. 2008]. Observa-se, então, que opiniões são importantes porque geralmente, antes das pessoas tomarem decisões, elas buscam opinião de outras. E isto não é somente verdade para indivíduos como também para organizações. No passado, quando uma pessoa queria opinião antes de comprar um carro, por exemplo, ela buscaria opinião de amigos e familiares. Olhando para as grandes organizações, se ela quisesse saber a opinião do público sobre seus produtos e serviços, ela realizaria pesquisas em grupos pequenos de pessoas.

Hoje em dia, com o crescimento explosivo das redes sociais e do conteúdo gerado na Web, essa realidade se transformou. As pessoas, por conta própria, publicam resenhas de produtos em sites comerciais, ou mesmo em seus próprios perfis nas mídias sociais. E o mais interessante é que as pessoas fazem questão de expor sua opinião quando estão em dois extremos de satisfação ou insatisfação com o produto. Esse efeito também ocorre em pessoas que estão envolvidas em serviços populares [LIU 2010.b].

A realidade então começa a mudar. Agora, se alguém quer adquirir novos serviços, ou produtos, não existe mais a limitação de somente procurar por opinião de amigos e familiares, pois vão existir até centenas de opiniões na Internet. Para uma empresa, a situação também muda, pois eles não precisam investir em pesquisas por grupo a fim de avaliar a satisfação com seus serviços: haverá uma abundância de tais informações disponíveis publicamente por clientes.

Dentro desse novo contexto, é possível ver que muitos usuários já buscam em sites de opinião, blogs ou fóruns, por opiniões antes de adquirirem seus produtos. Essa prática, contudo, não traz resultados muito satisfatórios, uma vez que o usuário precisaria avaliar uma quantidade muito grande de informação para obter uma avaliação mais concisa do que ele está buscando. Pode ser difícil e tedioso para um leitor humano comum encontrar sites relevantes, extrair frases relacionadas com opiniões, lê-los, resumi-los e organizá-los em formas utilizáveis.

2.4.1. Conceitos Básicos

Nesta seção são apresentados conceitos básicos para o entendimento de Análise de Sentimentos. Por ser uma área nova, não existe um padrão para os conceitos utilizados na área, de maneira que os termos utilizados não são universais.

Vamos usar o Exemplo 1, a seguir, para o entendimento geral dos conceitos:

> *"[1] Comprei um Samsung Galaxy S3 há alguns dias. [2] A qualidade do touchscreen é realmente boa, [3] e os recursos de câmera também são ótimos. [4] No entanto, o tempo de vida da bateria não foi bom para mim..."*.

Exemplo 1

A primeira questão a se pensar é se a passagem trata-se de um texto subjetivo ou opinativo. Sendo este um texto opinativo, então as opiniões devem ser extraídas e analisadas. Essa análise sobre texto opinativo e subjetivo será explicada na próxima seção.

Após observar rapidamente o texto, vê-se que as frases [2] e [3] classificam o objeto de maneira positiva, em relação aos aspectos *touchscreen* e *câmera*. Já na frase [4] temos uma opinião contrária ao que foi apresentado anteriormente, em relação ao aspecto *tempo de vida de bateria*. Percebe-se, portanto, que embora o objeto seja o mesmo (Samsung Galaxy S3), as características analisadas são diferentes.

Neste trabalho o padrão seguido será o proposto por [LIU 2010.2]. Sendo assim, são apresentadas as principais definições básicas:

- **Objeto:** Um objeto ***O*** é uma entidade que pode ser um produto, uma pessoa, as vias de uma cidade, um evento ou tópico. Todo objeto está associado a um par ***O: (T, A)***, onde ***T*** é uma hierarquia de componentes, subcomponentes e ***A*** é um conjunto de atributos que caracterizam ***O.*** No exemplo acima, o objeto seria "Samsung Galaxy S3".
- **Aspecto ou Característica:** Um aspecto/característica é um atributo, propriedade, parte ou componente do objeto. Aspectos podem ser classificados como *explícitos* ou *implícitos*. Observando o Exemplo 1, nota-se que câmera e bateria seriam aspectos do Objeto (Samsung Galaxy S3). É possível também notar uma hierarquia de aspectos, que será mais bem exemplificada na Figura 2.

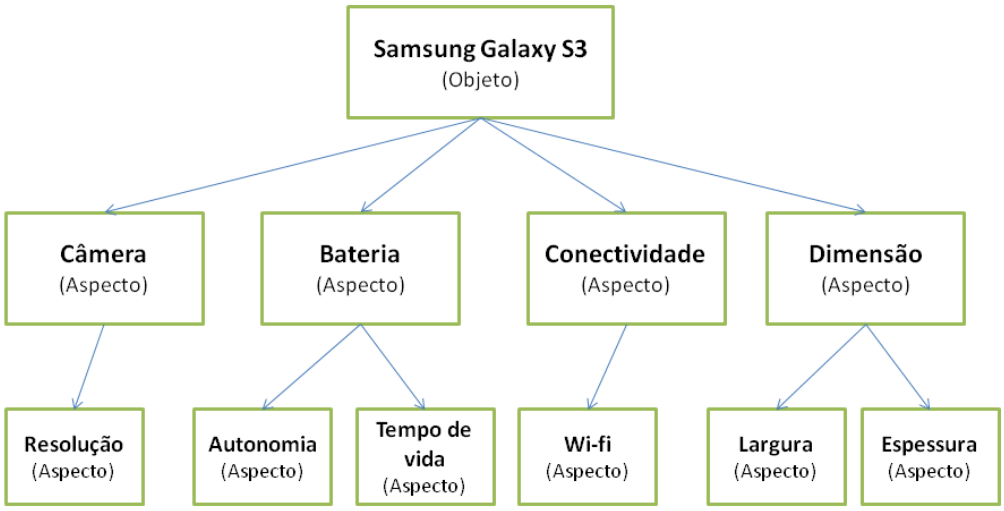


Figura 2 - Hierarquia Objeto X Características

- **Opinião:** uma *opinião* **demonstra uma visão, sentimento, atitude ou avaliação** sobre um objeto ou ainda sobre um aspecto desse objeto. Toda opinião tem uma orientação.
- **Orientação da Opinião:** A *orientação* de uma opinião sobre um objeto ou aspecto, **indica quando a opinião é *positiva*, *negativa* ou *neutr*a**. A *Orientação da Opinião* ainda pode ser chamada de *orientação semântica*, *orientação do sentimento* ou *polaridade da opinião*.
- **Palavras Opinativas:** São as palavras que qualificam os aspectos, sendo em sua maioria advérbios e adjetivos.

Outro conceito muito importante dentro do contexto de AS, é o de classificação dos documentos, que será detalhado nas etapas do processo de AS, na próxima seção.

2.4.2. Etapas da Análise de Sentimentos

Existe uma tendência sobre as etapas no processo geral de Análise de Sentimentos sugerida em [LIU 2010], que será descrita a seguir. A complexidade de cada uma das etapas pode ser tão alta que geralmente os trabalhos focam em uma ou duas dessas atividades. As etapas vão desde a coleta de dados até a sumarização dos resultados. As etapas sugeridas por [LIU 2010] são combinadas com as etapas sugeridas por [AGARWAL et al. 2011].

Coleta dos dados e Pré-processamento

A primeira fase consiste na *coleta de dados* onde devem ser buscados dados ou informações que estejam relacionados ao objeto de pesquisa. Após isso, existe uma etapa de pré-processamento que varia de um projeto para o outro. Geralmente é nessa etapa onde são excluídas palavras que não fazem sentido à análise e rotinas de "*tokenização*" dos documentos.

Análise da Subjetividade

Com os dados coletados, a segunda etapa da Análise de Sentimentos consiste em identificar a subjetividade no texto. Quando as frases possuem cunho subjetivo, ou seja, não apresentam uma opinião, elas são descartadas do processo de AS.

Segundo [PANG & LEE 2008], a tarefa de identificação da subjetividade nas sentenças de um texto pode se tornar mais complexa do que a classificação da opinião do texto. Existem técnicas que se preocupam com essa análise, como a de [AKKAYA & WIEBE - 2009] que busca usar a técnica de [WIEBE et al. 2005] para analisar a subjetividade de um texto, bem como a ambiguidade.

Extração de Características

Esta etapa é responsável por identificar os aspectos ou características que estão ligados ao objeto a ser analisado. É importante saber que ao se ter uma análise positiva em relação a um objeto, não significa que todas as características foram positivamente classificadas em relação ao objeto analisado. No entanto, pode significar que a maioria desses aspectos têm classificação positiva, podendo alguns ter sido negativamente classificados.

Observando o Exemplo 1, os aspectos extraídos ali seriam: *touchscreen*, *câmera* e *bateria*, pois esses revelam características que podem ser analisadas quando comparados dois ou mais celulares. Esses são os aspectos a serem classificados no texto.

A extração de características é uma tarefa difícil de ser realizada e automatizada no processo da Análise de Sentimento. Em muitos casos é preciso considerar o domínio, ou seja, utilizar uma técnica específica para um determinado domínio proposto, para que seja possível automatizar e extrair todas as características sem cometer erros [SIQUEIRA, 2010].

Há quatro técnicas para seleção de características que têm sido usadas nos estudos sobre AS. Esses métodos incluem extração de características sintáticas, semânticas, *link-based* e *estilísticas* [AGARWAL et. Al. 2011]. Como são muitas técnicas, serão apresentadas duas delas, a seguir, que são importantes para o entendimento do trabalho.

- **Técnica de extração de características semânticas:** incorpora técnicas de anotação manual/semiautomática ou totalmente automática para adicionar polaridade ou afetar a intensidade das características, podendo também relacionar pontuação para palavras e frases [AGARWAL et al. 2011].
- **Técnica manual/semiautomática de palavras geradas**: normalmente é usado um conjunto inicial de termos previamente gerados. Estes termos são codificados manualmente com a polaridade e a intensidade da informação [TONG 2011].

Classificação

Depois de identificar as características referentes ao domínio, o próximo passo será classificar a orientação da sentença. Essa classificação pode ser dividida em níveis, e diversas técnicas diferentes podem ser aplicadas para chegar a um melhor resultado da classificação, a depender do domínio.

A classificação pode ser dividida em níveis. Seguindo o padrão de [LIU 2010] e [FERNANDES, 2010], um documento pode ser classificado em nível de característica ou em nível de documento:

- **Classificação em Nível de Documento**: é quando a opinião é sobre todo o texto. A opinião é classificada como *negativa* ou *positiva* levando em conta todo o texto.
- **Classificação em Nível de Aspecto**: é quando a classificação não é no texto todo, mas sim acontece baseado em vários aspectos.

Se voltarmos para o Exemplo 1 podemos ver que se fosse feita uma classificação em nível de documento, provavelmente a classificação seria positiva, visto que existem duas sentenças positivas e uma negativa. No entanto, esse tipo de classificação perde muita informação, e pode ser ainda imprecisa. Já utilizando classificação em nível de aspecto, a opinião referente a cada característica será considerada.

Com isso em mente, é importante notar que as sentenças precisam ser classificadas, mas nem todas as frases são classificáveis [PANG & LEE 2008]. Existem as sentenças ***opinativas*** (Exemplo 1, frases 2, 3 e 4), onde é mostrada a opinião do autor com relação ao objeto ou aos aspectos. E ainda, existem as sentenças ***subjetivas*** (Exemplo 1, Frase 1), onde as frases não apresentam opinião, e sim fatos sobre determinado objeto.

Para realizar a classificação, vários passos são realizados, utilizando-se de diversas técnicas. Abaixo são descritos os passos de um algoritmo de classificação [DING et al. 2008]:

- **Extrair e classificar as palavras opinativas**: nesse primeiro passo, o objetivo é encontrar as palavras opinativas contidas na sentença, e classificá-las como positivas, negativas ou neutras. Retornando ao Exemplo 1 novamente, as sentenças podem ser classificadas como: "*A qualidade do touchscreen é realmente boa **[positivo]**, e os recursos de câmera também são ótimos [**positivo**]. No entanto, o tempo de vida da bateria não foi bom para mim **[positivo]***".
- **Cláusulas Negativas:** Um segundo passo é identificar as cláusulas negativas, que irão inverter a polaridade das palavras opinativas as quais ela se refere. Na terceira sentença do Exemplo 1, apesar de a palavra opinativa "***bom"*** ser classificada como positiva, a sua polaridade é invertida por causa da cláusula negativa "***não"*** que a antecede. Logo, a sua classificação iria ficar da seguinte forma: "*No entanto, o tempo de vida da bateria não foi bom para mim **[negativo]***".

2.4.3. Desafios e Limitações

Análise de Sentimentos é uma subárea de Processamento de Linguagem Natural bastante recente, o que faz com que ainda exista muitos desafios e limitações a serem superados. Nas próximas seções, serão citados alguns dos desafios enfrentados na AS.

Classificação da Polaridade dos Adjetivos e dos Advérbios

Classificar a polaridade de um adjetivo não é apenas uma tarefa complexa, como também existem poucos resultados na literatura. Um dos pontos iniciais para se classificar um texto é a classificação dos adjetivos e advérbios. Em muitos trabalhos, essa classificação é feita pelo próprio usuário [FERNANDES, 2010], ou ainda muito dependente do domínio da análise.

2.5. Análise de Sentimentos e Trânsito

De acordo com a Seção 2.3, é possível resumir que a área de análise de sentimentos se preocupa em extrair o sentimento dentro de textos opinativos.

Com o crescimento do uso de Redes Sociais como meio de comunicação, especialmente *microblogs* como *Twitter*, tem sido cada vez mais comum no Brasil o compartilhamento de informações de trânsito nas Redes Sociais.

A partir desse estudo, entende-se que é possível extrair um *sentimento* de mensagens sobre o trânsito em geral. A seguir, vamos detalhar um *tweet* comum sobre "trânsito" e mostrar os elementos inerentes à Análise de Sentimentos.

> *"Trânsito completamente parado pra quem vai pela Avenida Beberibe".*

Exemplo 2 - Fonte: twitter.com/JcTransito

No Exemplo 2, a sentença "*Avenida Beberibe*" pode ser claramente considerada como sendo o *Objeto* da frase, enquanto "*parado"* é uma característica do *Trânsito* (o domínio em questão). Ainda, "*completamente"* é uma palavra opinativa sobre a característica.

Portanto, a partir dessa relação, percebeu-se que as informações sobre o trânsito podem ser consideradas textos opinativos, já que expõem situações *positivas* ou *negativas* em relação ao trânsito.

2.6. Considerações

Com os conceitos apresentados, é possível concluir que o processo de Análise de Sentimentos é complexo devido à quantidade de etapas que o compõem. Por essa razão, em muitos trabalhos, algumas dessas etapas são realizadas de forma manual. No entanto, mesmo com a complexidade percebida, é possível construir sistemas e aplicações úteis para o cotidiano ou outros grandes sistemas.

Com esses conceitos fundamentados, no próximo capítulo será apresentado um estudo sobre os sistemas existentes que tratam Análise de Sentimentos considerando as mensagens de redes sociais.

3. Trabalhos Relacionados

Este capítulo analisa e discute dois trabalhos que envolvem Análise de Sentimentos voltada a dados de *microblogs*. O primeiro artigo discutido é o trabalho de [AGARWAL et al. 2011]. Em seguida, é apresentado o estudo presente em [PANDEY & IYER 2010].

3.1. Sentiment Analysis of Twitter Data

O trabalho apresentado por [AGARWAL et. Al. 2011] consiste em examinar sentimentos contidos em dados do *Twitter* provenientes de uma fonte comercial (não especificada). O principal objetivo desse trabalho é construir modelos para **classificar** *tweets* como positivos, negativos ou neutros. Foram construídos três modelos:

- ***Unigram model*:** modelo baseado em busca de palavras únicas dentro do *tweet*, a fim de extrair o sentimento a partir dessas palavras únicas.
- ***Feature based model*:** após criar um dicionário de características acerca de um determinado domínio, o *feature based model* usa características criadas pelos próprios autores para agregar sentimento aos tweets.
- ***Tree kernel based model*:** nesse modelo foi desenvolvida uma árvore que representa os *tweets.* Esse modelo combina várias categorias de características em um só tipo de representação conveniente.

Destacamos no trabalho a fase inicial de pré-processamento e o modelo construído junto à seleção de características. As maiores vantagens percebidas são:

- Os autores se preocupam em analisar diferenças de palavras como “muuuuuito” e “muito”, onde cada uma tem uma participação diferente no cálculo da análise.
- São levados em consideração termos chamados *emotions*, que são expressões do tipo “=)” e “=(“. Esse tipo de tratamento agrega muito valor à classificação, o que pode se estender para diversos domínios de análise.

Contudo, na fase inicial, foi notada uma desvantagem do sistema: o conjunto de *tweets* selecionado para análise é “pré-anotado” por um humano. Na etapa de pré-seleção, os *tweets* coletados passam por uma fase de classificação que consiste em: *positivo, negativo, neutro* e *lixo.* Todos os dados classificados manualmente como *lixo* são descartados. Esse *lixo* consiste em mensagens mal traduzidas para o inglês ou que não façam parte do domínio. Os outros resultados são usados na análise final.

Já na análise de resultados, podemos perceber a eficácia do sistema. Dada a sua complexidade, o trabalho se preocupou em analisar resultados de classificação nos três modelos apresentados anteriormente. Num conjunto de cerca de cinco mil *tweets* – que não continham *lixo*, a taxa de acerto na classificação – comparada à classificação humana inicial – foi de aproximadamente 61%.

3.2. Sentiment Analysis of Microblogs

Esse estudo apresenta um sistema que coleta dados do *Twitter*, analisa as mensagens a partir do *sentimento* presente nelas classificando-as como neutras, positivas ou negativas. Nesse trabalho foram testados diferentes modelos de *extração de características*, bem como algoritmos de classificação, buscando uma melhor combinação dos dois modelos.

Para a classificação das mensagens, a abordagem usada em [PANDEY & IYER 2010] lida com dois classificadores que separa as mensagens nas três categorias desejadas. O primeiro classificador, *Neutral-Polar Classifier*, separa as mensagens em *neutras* e *polares* (negativas ou positivas). As mensagens consideradas *polares* por esse primeiro classificador são passadas para um segundo classificador, *Positive-Negative Classifier*, que foi treinado para distinguir *sentimentos positivos* dos *negativos.* A Figura 3 mostra os componentes desse sistema.

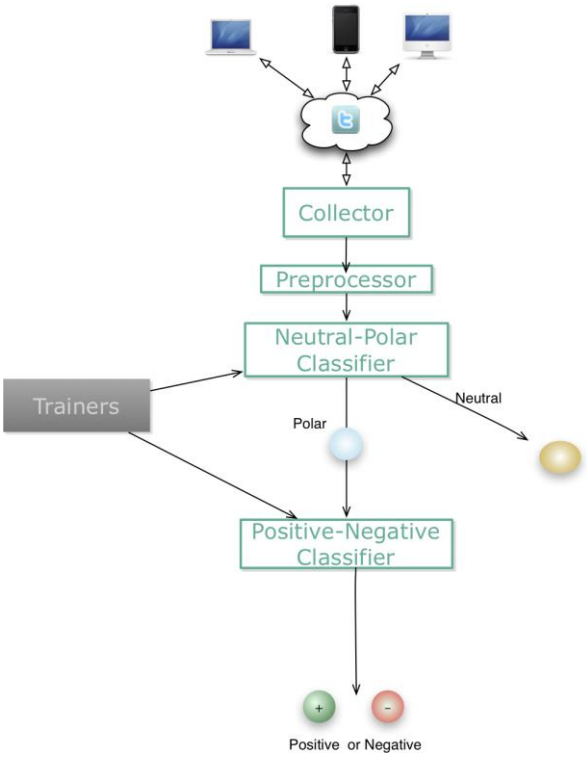


Figura 3 - Componentes do sistema [PANDEY & IYER 2010].

As primeiras etapas seguidas no desenvolvimento desse sistema são as tradicionais, como sugeridas em [AGARWAL et al. 2011] e [LIU 2010]: Coleta dos Dados, Pré-Processamento, Extração de Características.

Após a etapa de Extração de Características, a fase seguinte é particular desse trabalho. Chamada de *Model Selection,* esta é a etapa onde um modelo é treinado, usando conjuntos de treinamentos pré-anotados manualmente. O desenvolvimento é, então, seguido pela etapa de Classificação, onde o modelo treinado classifica as novas mensagens.

Foram coletados mais de vinte mil *tweets* mencionando o produto pertencente ao domínio do sistema (não especificado). Assim como [AGARWAL et al. 2011], os dados iniciais também foram pré-classificados por um humano. No entanto, no estudo de [PANDEY & IYER 2010], as mensagens foram "pré-anotadas" como *positivas*, *levemente positivas, negativas, levemente negativas* ou *neutras.* Essas mensagens são usadas para treinar o *Positive-Negative Classifier*.

Ao longo desse estudo são detalhadas as formas de pré-processamento, os algoritmos usados nos classificadores bem como os diferentes resultados obtidos com os modelos de seleção (*Model Selection*).

Chamamos, então, atenção para os resultados obtidos. Os autores afirmam que o resultado do *Neutral-Polar Classifier* não foi tão satisfatório, pois muitas mensagens neutras acabam passando para o *Positive-Negative Classifier*, dado um alto número de falsos positivos.

O trabalho teve problemas com mensagens ambíguas ou irônicas. O fato de retirá-las manualmente é uma desvantagem, pois afeta no resultado final de avaliação do desempenho do modelo de classificação. Os autores também afirmam que treinaram mais de dois mil modelos de classificadores para chegar a um resultado satisfatório, o que se considera custoso.

3.3 Outros trabalhos relacionados

Além dos dois trabalhos detalhados, é importante levar em consideração outros estudos que envolvem técnicas de AS relacionadas ao *Twitter*, mas com uma abordagem diferente.

Em [BIGONHA ET al. 2010], por exemplo, o objetivo é detectar perfis do *Twitter*, ou pessoas influentes a partir de análise de sentimentos. Nesse estudo são usadas não só técnicas de descoberta de polaridade, como também rede de contatos, citações e outros recursos do *Twitter* que podem ser explorados.

Já o estudo de [BIFET & FRANK 2010] consiste em minerar dados do *Twitter* em tempo real e analisar o desempenho desse processo. O trabalho não usa AS, mas é envolvido com a análise de *tweets* e possui etapas de pré-processamento comuns às apresentadas.

Outro artigo que se destacou foi o de [CUI et al. 2011], cujo objetivo é analisar *tokens* que carregam *sentimentos*. Ou seja, palavras consideradas como *características* são analisadas, a fim de distinguir o significado e aumentar o nível de precisão de sua polaridade.

3.4 Considerações

Após analisar vários trabalhos, destacamos a semelhança entre suas etapas. A grande maioria dos trabalhos analisados seguem as etapas do processo de AS definido no estudo de [LIU 2010], [AGARWAL et al. 2011] e [PANG & LEE 2008]. Isso mostra a coerência que a área de Análise de Sentimentos vem construindo nos últimos anos, bem como importância que esses estudos consolidativos têm para os novos trabalhos.

No entanto, grande parte dos trabalhos na área envolve análise de mensagens voltada à avaliação de produtos, serviços ou empresas. Isso deixa espaço para que novas aplicações possam surgir, com o intuito de analisar outros nichos, como por exemplo: trânsito, meios de transporte, reputação de pessoas, entre outros.

4. UbibusAnalysis

Este é o capítulo que descreve a ferramenta desenvolvida. Será mostrado o que foi compreendido como solução em Análise de Sentimento para o problema tratado, a arquitetura geral e os detalhes sobre a implementação do UbibusAnalysis.

4.1. O que é o UbibusAnalysis?

O UbibusAnalysis foi criado para ser uma ferramenta de análise de mensagens de trânsito para o sistema Ubibus [VIEIRA et al. 2011]. Em nosso trabalho, a abordagem escolhida para analisar as mensagens de trânsito foi com Análise de Sentimentos (AS). A proposta do trabalho é construir uma ferramenta de análise completa, que trabalhe desde a coleta de mensagens nas Redes Sociais, as etapas de análise do texto, até a entrega dos resultados para um usuário final ou uma aplicação.

4.2. Arquitetura

A arquitetura é dividida em vários componentes, são eles: base de dados, buscador de mensagens, analisador de sentimentos e API Rest. Os componentes e seus subcomponentes serão detalhados a seguir.

O TweetsCrawler guarda as mensagens no formato original na Base de Dados, o UbibusAnalyzer trata as mensagens, cria ocorrências, e volta a armazená-las. O AnalysisAPI provê os serviços. A arquitetura geral pode ser vista na Figura 4, e os detalhes se encontram nas próximas seções.

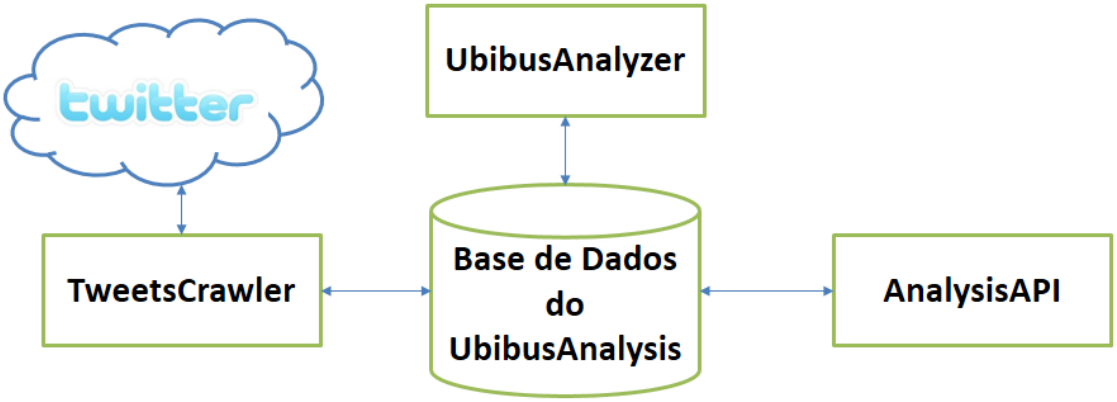


Figura 4 - Arquitetura Geral do UbibusAnalysis

4.2.1. Base de dados do UbibusAnalysis

O UbibusAnalysis foi projetado para funcionar seguindo o padrão da Base de Dados do Ubibus (**Apêndice 1**). Como o Ubibus [VIEIRA et al. 2011] é um sistema que aglomera diversas aplicações, apenas parte de sua base de dados é utilizada no trabalho, como pode ser visto na Figura 5.

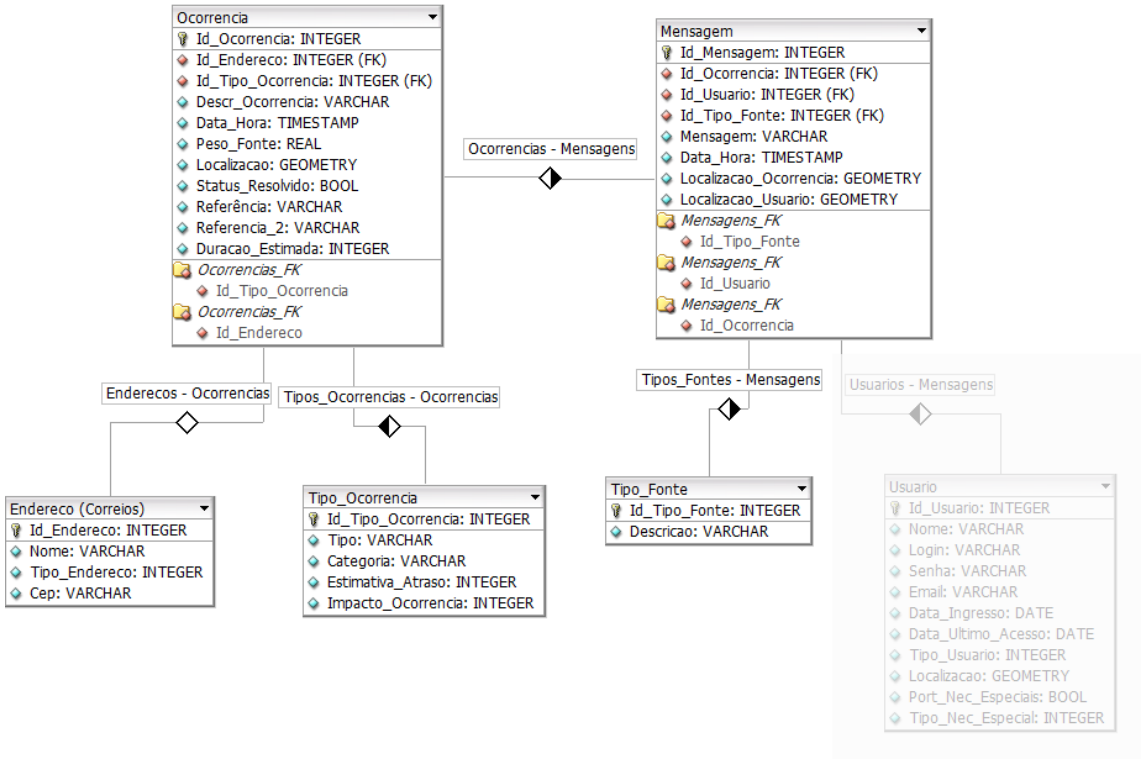


Figura 5 - Entidades usadas no UbibusAnalysis

As principais entidades utilizadas são: **Mensagem** e **Ocorrência**, pois pode-se observar as **Mensagens** como sendo a fonte da análise e **Ocorrência** como sendo o resultado.

- **Endereço:** entidade que representa os endereços. Composta por um tipo, CEP e o nome completo.
- **Tipo Fonte:** são os tipos das fontes das mensagens. Identificam Redes Sociais x Perfis dessas redes.
- **Tipo Ocorrência:** caracteriza como uma ocorrência pode ser interpretada, seu impacto e suas características quanto à positividade ou negatividade. O **Apêndice 2** mostra detalhes sobre essa tabela.
- **Mensagem:** toda mensagem possui dados de sua fonte, data e hora, localização. Além disso, toda mensagem deve ser associada a uma ocorrência. No UbibusAnalysis, essa associação ocorre após análise da mensagem.
- **Ocorrência:** é uma interpretação do que uma mensagem pode trazer. Está associada a uma localização e a um tipo de ocorrência.
- **Usuário:** é uma entidade associada a uma mensagem mas não é usada no trabalho. Todas as suas referências são nulas.

4.2.2. Buscador de mensagens: TweetsCrawler

TweetsCrawler é o componente que busca e coleta mensagens em Redes Sociais. Como as mensagens abordadas no trabalho são curtas, a fonte desses dados foi o *Twitter*. Devido ao alto volume de dados no *Twitter*, e sua característica dinâmica, o TweetsCrawler funciona de forma sistemática (Figura 6).

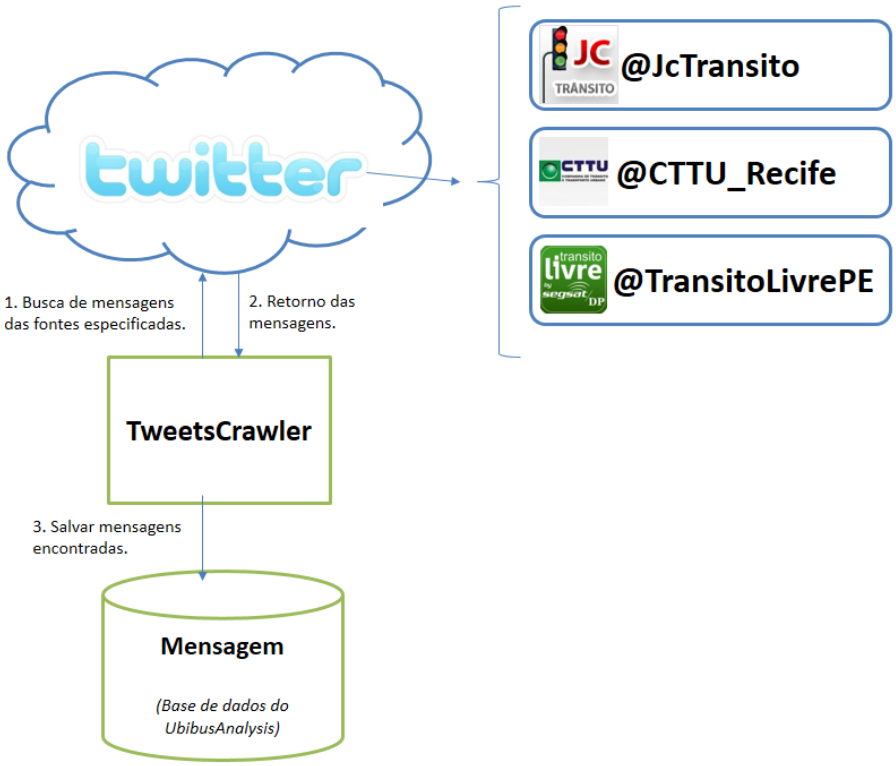


Figura 6 - TweetsCrawler

Este buscador funciona de forma independente e periódica. **A cada cinco minutos**, o TweetsCrawler busca por mensagens de trânsito da cidade de Recife das fontes @JcTransito[1], @CTTU_Recife[2] e @TransitoLivrePE[3]. Caso sejam encontradas mensagens que ainda não foram salvas na base de dados, na tabela **Mensagem**, o buscador se encarrega de salvá-las. Como as mensagens não são analisadas por esse componente, a referência para **Ocorrência** é nula. Salvar uma **Ocorrência** é um dos objetivos do Analisador de Sentimentos, descrito na seção a seguir.

4.2.3. Analisador de Sentimentos: UbibusAnalyzer

O objetivo do Analisador de Sentimentos é criar ocorrências a partir de mensagens. Para isso, esta seção explica qual o processo executado pelo UbibusAnalyzer.

Na tabela **Mensagem**, descrita na Figura 5, observamos que existe uma referência para a tabela de **Ocorrência**. A rotina de coleta de mensagens não ocorre simultaneamente à de análise. Os dois componentes, TweetsCrawler e UbibusAnalyzer possuem rotinas independentes, embora os dados possuam dependência. A arquitetura do UbibusAnalyzer, e os componentes com o qual ele se comunica, é descrita na Figura 7.

[1] @JCTransito <http://twitter.com/JcTransito>
[2] @CTTU_Recife <http://twitter.com/CTTU_Recife>
[3] @TransitoLivrePE <http://twitter.com/TransitoLivrePE>

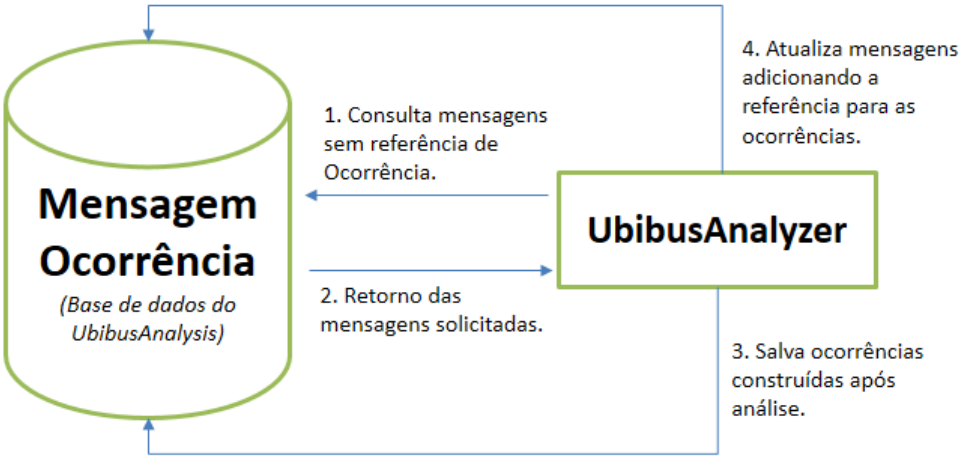


Figura 7 - UbibusAnalyzer e componentes adjacentes

A cada intervalo de tempo, definido como dez minutos, o **UbibusAnalyzer** inicia seu procedimento. É realizada, então, uma consulta à base de dados que consiste em selecionar todas as mensagens que ainda não possuem uma referência para alguma ocorrência. Após ter esse conjunto de mensagens, que será chamado de *conjunto para análise*, o Analisador executará um processo de análise em particular para cada uma das mensagens contidas nesse conjunto.

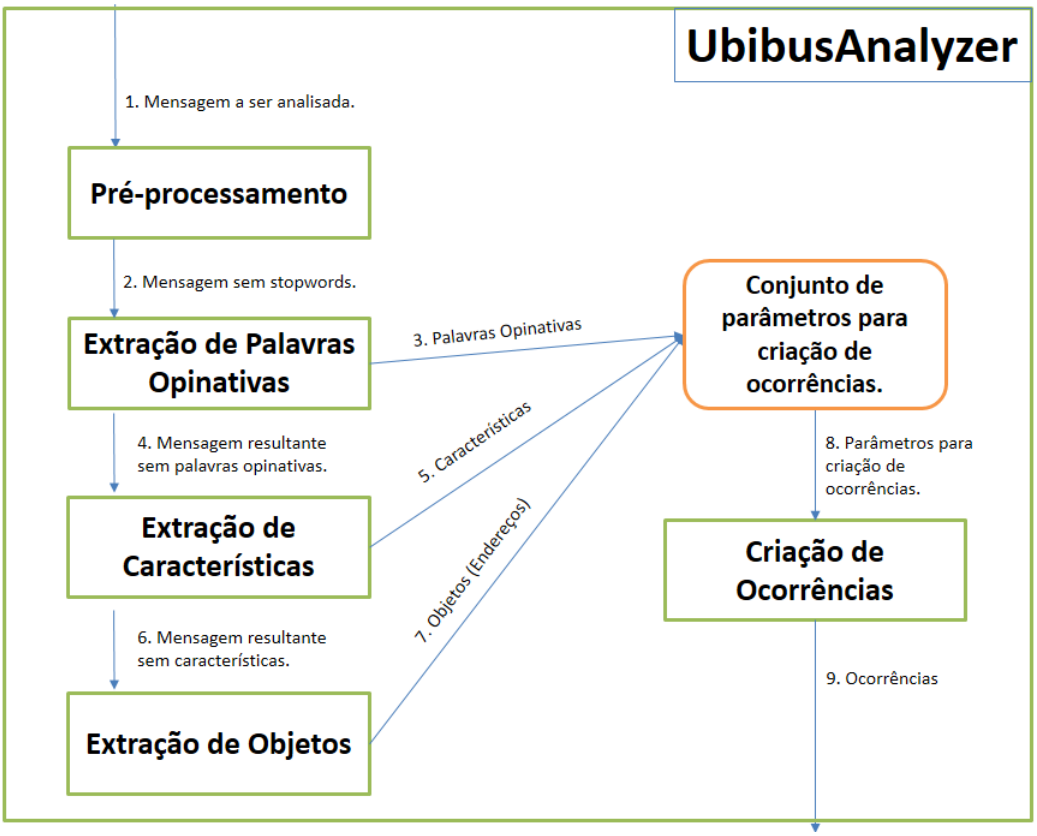


Figura 8 - Análise de Sentimentos por mensagem

A partir daí, o analisador tenta extrair palavras opinativas, características e uma localidade, que serão reunidos ao final como parâmetros para a construção da(s) ocorrência(s). O processo é apresentado na Figura 8.

Caso não existam mensagens no *conjunto para análise*, o UbibusAnalyzer fica em estado de espera até completar o intervalo de tempo de cada período e buscar por novos *tweets*.

Pré-processamento

A primeira etapa da análise é a de pré-processamento. No UbibusAnalyzer, usamos o conceito de *stopwords.* São consideradas *stopwords* todo tipo de caractere ou palabra que não influencia no processo da análise. A camada de pré-processamento do UbibusAnalyzer é dividida da seguinte forma:

1. Primeiramente são removidas as stopwords do tipo pontuação e caracteres especiais, encontradas no **Apêndice 3**.
2. Retirados os caracteres do tipo pontuação, a segunda etapa é retirar da mensagem as *stopwords* de pré-processamento (**Apêndice 4**). Essas são palavras que não influenciam na análise posterior. Em sua maioria, são fora do domínio da aplicação, que é o *trânsito,* além de preposições, artigos, conjunções e pronomes. Nesse conjunto, também estão contidas palavras do tipo menção de *Twitter* (@mentions), *URLs*, e palavras em formato de hora (por exemplo, 12h30min, 14h10, 8:05, entre outras).

Extração de Palavras Opinativas

Após a camada de pré-processamento, a aplicação segue para uma camada chamada **Extração de Palavras Opinativas**. O objetivo dessa fase é coletar as palavras opinativas, guardá-las em um conjunto e retirá-las da mensagem. Dessa forma, o texto da mensagem chega mais "enxuto", facilitando as próximas etapas. O dicionário de palavras opinativas se encontra no **Apêndice 6**. Lá elas são especificadas junto às suas categorias (positivas ou negativas).

Extração de Características

A camada seguinte consiste na **Extração de Características**. Essa etapa é fundamental porque se não forem encontradas características na mensagem, o UbibusAnalyzer considera a mensagem como *sem polaridade*, o que então não caracteriza uma ocorrência de trânsito. Quando encontradas, as características são incluídas em um conjunto à parte e retiradas da mensagem. No **Apêndice 7** são descritas as características de acordo com seu tipo, sua categoria e seu impacto (detalhes necessários na fase de implementação).

Com a mensagem sem *características* e sem *palavras opinativas*, aplicou-se novamente uma remoção de *stopwords*, contida no **Apêndice 5**. Durante o desenvolvimento da aplicação, verificou-se a necessidade dessa fase intermediária, **devido a algum *lixo* que a primeira fase não conseguiu retirar da mensagem**. São em geral preposições e artigos. Várias dessas palavras foram incluídas nesse dicionário graças à observação das fontes das mensagens.

Extração de Objetos: a descoberta dos endereços

A etapa que conclui essa fase do UbibusAnalyzer é a extração de objetos. Como nessas mensagens, o Objeto são os endereços onde acontecem as ocorrências, a última etapa cuida

dessa parte, já que a mensagem recebida já se encontra sem *lixo*. Na maioria dos casos, como será avaliado no próximo capítulo, a mensagem só chega nessa etapa praticamente com os endereços.

Extraídos, então, os endereços, eles são passados para um conjunto de objetos, e encaminhados para a próxima fase do UbibusAnalyzer. Chamamos a observação para esse ponto. O *conjunto para análise* de mensagens, são as mensagens a ser analisada. Já o resultado obtido após as fases iniciais do UbibusAnalyzer: extração de palavras opinativas, características e objetos, consistem na formação de novos conjuntos. São detalhados no próximo tópico.

Criação de uma Ocorrência

Ao chegar nessa fase, esperam-se os conjuntos do que foi extraído (e que é necessário) para a criação da ocorrência. Neste ponto do analisador, teremos:

- *Conjunto de palavras opinativas*, contém todas as palavras opinativas e suas respectivas categorias.
- *Conjunto de características*, contém todas as características encontradas na mensagem. Caso não existam características, não é possível criar um objeto do tipo **Ocorrência**. Nessa situação, essa mensagem tem como referência uma ocorrência neutra, que indica que essa mensagem não contém informação de trânsito.
- *Conjunto de objetos*, contém um ou mais endereços encontrados na mensagem. Para cada endereço é criada uma ocorrência. No entanto, a mensagem só referenciará uma delas.

Com esses conjuntos em mãos, foi criado um algoritmo que calcula a pontuação de cada característica junto com as palavras opinativas, e então é criada uma Ocorrência.

4.2.4. API Rest do UbibusAnalysis: AnalysisAPI

API, de *Application Programming Interface* (ou Interface de Programação de Aplicativos) é um conjunto de rotinas e padrões estabelecidos por um software para a utilização das suas funcionalidades por aplicativos que não pretendem envolver-se em detalhes da implementação do software, mas apenas usar seus serviços[4].

Já REST[5] significa Transferência de Estado Representacional, do inglês *Representational state transfer.* É uma técnica de engenharia de software para fornecer uma camada de comunicação em cima do protocolo de internet HTTP. REST não é um protocolo, não possui padrão, mas qualquer pessoa pode fornecer a API REST do seu site.

Isso só é possível, porque REST trabalha apenas com requisições de URL no *browser*, logo, qualquer dispositivo que acesse a internet, poderá enviar requisições e receber respostas de um site utilizando a API REST dele. Com esta ferramenta, é possível fazer a abstração do dispositivo que está fazendo a requisição, podendo ele ser um *desktop*, um *tablet* ou um celular.

[4] **FOLDOC** <http://foldoc.org/Application+Program+Interface>
[5] **REST** <http://rest.elkstein.org/>

Por esse motivo, o UbibusAnalysis disponibiliza, inicialmente, dois serviços em uma API seguindo os padrões REST, só assim os dados da análise podem ser consumidos facilmente.

Serviços oferecidos

O primeiro serviço é o de **busca de situação de endereço por horário e data**. Na URL de requisição, são necessários parâmetros de endereço, data e um intervalo de tempo. O resultado consiste nas ocorrências que ocorrem nesse espaço de tempo.

O parâmetro *intervalo de tempo* serve para indicar a variação máxima do horário da ocorrência em relação ao horário passado como parâmetro. Por exemplo, se um usuário requisita a situação do endereço **Avenida Caxangá** às **11:30** do dia 30/03/2013, num intervalo de **1 hora**, então todas as ocorrências entre **10:30** (uma hora antes) e **12:30** (uma hora depois) são retornadas. Caso não existam ocorrências registradas nesse horário, ou para esse endereço, um conjunto vazio é retornado como resposta.

O segundo serviço oferecido é o de categorização de mensagens. Para isso, o usuário precisaria, na URL de requisição, passar uma mensagem como parâmetro. O resultado consistiria das características extraídas da mensagem e se ela é positiva ou negativa.

5. Implementação e Resultados obtidos

Após tomar conhecimento do UbibusAnalysis e seus componentes, além de como eles interagem, este capítulo irá mostrar detalhes de como cada um desses componentes foram implementados, algoritmos utilizados, entre outros. Além disso, são mostrados na segunda seção detalhes sobre os resultados obtidos e alguns desafios.

5.1. Implementação

5.1.1. Base de dados do UbibusAnalysis

A base de dados do UbibusAnalysis foi construída utilizando o modelo relacional do banco PostgreSQL. Apesar disso, o UbibusAnalysis usa Django[6]: um *framework* web e *open source* para construção rápida de aplicações e servidores que oferecem serviços Web.

Com o uso do Django não foi necessário se preocupar tanto com a modelagem da base de dados do do Ubibus [VIEIRA et al. 2011], pois o Django transforma todas as entidades para *modelos* Python[7]. E então as entidades passam a ser enxergadas como objetos Python. Por esse motivo, a linguagem de programação escolhida para o desenvolvimento de todos os componentes foi Python.

5.1.2. TweetsCrawler

O TweetsCrawler busca informações do *Twitter* a partir da API Rest disponibilizada pelo *Twitter*. O componente realiza requisições URL passando como parâmetro o nome do perfil desejado, a quantidade de t*weets* a ser recebida e mais alguns detalhes, como: receber *Tweets* que são do tipo *retweet*, ou *mention.*

No UbibusAnalysis, a busca que ocorre num intervalo periódico de cinco minutos, busca por mensagens que não sejam do tipo *retweet* e nem do tipo *mention*, o que reduz o número de mensagens que podem não conter ocorrências de trânsito.

Ao observar a arquitetura, descrita no capítulo anterior (Figura 6), quando o TweetsCrawler faz uma requisição, ele a faz referente a um perfil por vez, coleta os *tweets* dos últimos cinco minutos referente àquela fonte, e salva as mensagens no formato da Base de Dados (Figura 5). Após isso, o componente busca as mensagens da próxima fonte, até que se esgotem as fontes. E então aguarda por um novo período de cinco minutos.

Para evitar trabalho duplicado, caso alguma mensagem com o mesmo texto já esteja na base de dados, ela não é salva. Já que seriam extraídas as mesmas ocorrências dessas mensagens.

5.1.3. UbibusAnalyzer

O UbibusAnalyzer é o componente responsável por analisar as mensagens. No capítulo anterior, foram apresentadas cinco etapas no processo de análise. Todas elas buscam por palavras-chave dentro do texto de uma mensagem. Essas palavras são então armazenadas num conjunto à

[6] **Django** <https://www.djangoproject.com/>
[7] **Python** <http://www.python.org>

parte e removidas da mensagem original. O processo de busca é o mesmo para as quatro primeiras etapas da análise de sentimentos.

Busca de palavras dentro de uma mensagem

Dada uma mensagem, o que o componente de extração de palavras faz é separar a mensagem em *tokens* individuais. Cada *token* é uma *string* (cadeia de caracteres) da mensagem original. Após isso, o buscador percorre a lista de *tokens* buscando uma *string* **similar** à palavra passada como parâmetro. Caso a palavra seja similar, então ela é inserida no conjunto correspondente e retirada dessa lista de *tokens.*

Ao final do processo, o buscador reúne os *tokens* numa mensagem (separados por espaço em branco), as palavras removidas em um conjunto e retorna esse resultado.

Similaridade entre palavras

Como mencionado no tópico anterior, o **buscador** procura por *strings* similares. Essa similaridade é calculada por um algoritmo da biblioteca **Difflib**[8] de Python. Quando enviadas para a função que calcula a similaridade, as duas *strings* são normalizadas, isto é, têm seus caracteres no formato minúsculo e sem acentos. A ilustração desse processo pode ser vista na Figura 9.

O algoritmo de **Difflib** calcula a similaridade entre duas palavras de acordo com o número de caracteres em comum e sua posição na *string*. De posse dessas duas informações, o algoritmo calcula uma média entre 0 e 1. O valor 0 (zero) representa nenhuma similaridade, enquanto o valor 1 indica que as duas *strings* são idênticas.

Nas etapas de pré-processamento, extração de palavras opinativas e extração de características, consideramos duas palavras como similares, *strings* cuja similaridade calculada pelo algoritmo Difflib fosse igual ou superior a 0.8 (80% de similaridade).

Na etapa de extração de objetos, como as cadeias comparadas são compostas por mais de uma palavra, então o nível de similaridade aceito foi de 0.75 (75%). A comparação acontece entre a cadeia resultante e o nome dos endereços armazenados na Base de Dados do UbibusAnalysis.

De acordo com testes realizados manualmente, os valores de 0.8 e 0.75 de similaridade foram considerados aceitáveis. Isso porque, no primeiro caso, os *tweets* podem conter erros de digitação. Com essa tolerância, dados que podem ser relevantes não são perdidos.

Já no caso de extração de objetos, pode acontecer de os endereços não serem digitados com seu nome completo. Por exemplo, o endereço “Avenida Caxangá” muitas vezes é simplesmente chamado de “Caxangá”. Para dar suporte a situações como essa, o valor de 0.75 apresentou resultados satisfatórios. Mais detalhes sobre os resultados são avaliados no final deste capítulo.

[8] **Difflib – Documentação oficial** <http://docs.python.org/2/library/difflib.html>

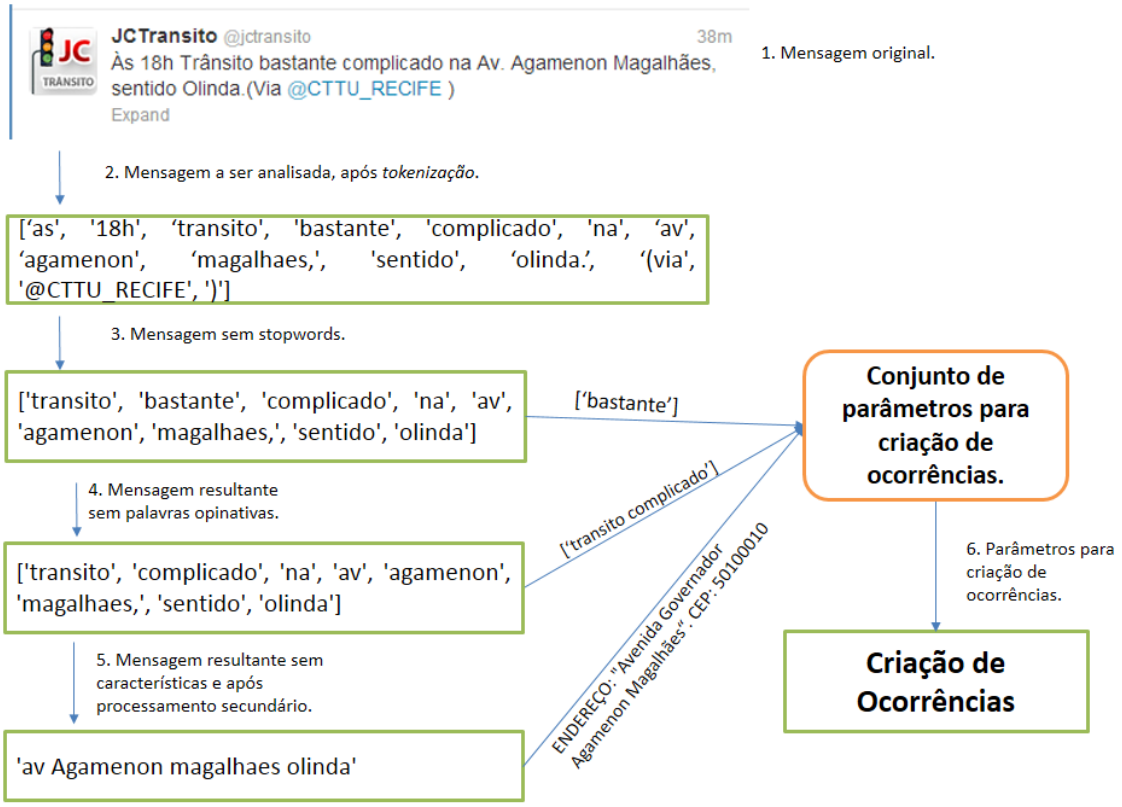


Figura 9 - Análise de mensagem

Criação de Ocorrência e cálculo de Impacto

Observando a estrutura da entidade Ocorrência, descrita na Figura 5, vemos que ela possui três campos essenciais: um tipo de ocorrência, um endereço e um impacto.

O impacto da ocorrência é fundamental para a escolha do tipo de ocorrência. Na criação do **Apêndice 7**, para cada característica, foram adicionados os atributos **impacto_característica, impacto_mínimo, impacto_máximo** e **tipo.** O **impacto_caracteristica** significa o valor de impacto que o *tipo de ocorrência* relacionado àquela ocorrência assume.

Caso existam palavras opinativas, o impacto tem seu valor mudado. Como os valores trabalhados na base do Ubibus vão de 1 (zero) a 4 (quatro), onde 1 (um) e 2 (dois) são negativos e 3 (três) e 4 (quatro) são positivos, **mínimo** e **máximo** determinam a faixa onde um impacto daquele tipo de característica pode estar. Quanto menor o valor numérico do impacto, maior ele é, ou seja, mais ele afeta o trânsito.

A lógica de cálculo de um impacto, portanto, leva em consideração as palavras opinativas e o impacto das características. Palavras opinativas **positivas** enfatizam o *poder* de uma característica. Por exemplo, "lento" tem uma conotação, e "**muito** lento" significa que a característica "lento" já possui um impacto mais forte. Por outro lado, palavras opinativas **negativas** atenuam o *poder* de uma palavra opinativa, ou seja, "lento" seria mais forte que "**pouco** lento".

É importante, contudo, perceber que palavras opinativas negativas junto a características negativas podem demonstrar situações inversas. A lógica é detalhada na descrição do algoritmo de cálculo a seguir, que descreve como chegamos ao resultado do impacto da ocorrência (**IMPACTO_OCORRENCIA**) em relação ao impacto da característica (**IMPACTO_CARACTERISTICA**).

1. *Se não for encontrada PALAVRA_OPINATIVA, então: IMPACTO_OCORRENCIA = IMPACTO_CARACTERISTICA.*
2. *Se for encontrada palavra opinativa **positiva** com característica **positiva:** IMPACTO_OCORRENCIA = IMPACTO_CARACTERISTICA + 1.*
   Ou seja, numa situação positiva, com uma palavra opinativa positiva, a ocorrência se torna "melhor", com um valor numérico maior.
3. *Se for encontrada palavra opinativa **positiva** com característica **negativa:** IMPACTO_OCORRENCIA = IMPACTO_CARACTERISTICA – 1.*
   Nessa situação, a palavra **positiva** enfatiza a negatividade. Na frase "O trânsito se encontra muito lento". É o caso descrito anteriormente.
4. *Se for encontrada palavra opinativa **negativa** com característica **positiva**: IMPACTO_OCORRENCIA = IMPACTO_CARACTERISTICA – 1.*
   Nesse caso, numa situação positiva, com uma palavra opinativa negativa, a ocorrência se torna "pior", com um valor numérico maior. Por exemplo: "O trânsito está nada bom." **Nada** é negativa enquanto bom é positiva.
5. Se for encontrada palavra opinativa **negativa** com característica **negativa**: IMPACTO_OCORRENCIA = IMPACTO_CARACTERISTICA + 1.
   É o caso em que a palavra opinativa diminui a *força* de uma característica. É um exemplo de "Agora o trânsito já está menos devagar". **Menos** atenua a intensidade da palavra **devagar.**
6. Ao final do cálculo, se o IMPACTO_OCORRENCIA > 2, então o TIPO_OCORRENCIA = positivo. Se IMPACTO_OCORRENCIA < 3, então o TIPO_OCORRENCIA = negativo.

Após as características extraídas, os objetos resultantes são semelhantes aos do **Apêndice 7**. Para encontrar o **tipo_ocorrência** correto, a consulta então busca no banco de dados por um tipo de ocorrência que contenha o tipo daquela característica e o impacto calculado. Seria uma consulta semelhante à do Exemplo 3 - Consulta de Tipo de Ocorrência.

```
SELECT      T
FROM        TIPO_OCORRENCIA T
WHERE       t.tipo = <tipo_ocorrencia_extraido>
AND         t.impacto = <impacto_ocorrencia>
```

Exemplo 3 - Consulta de Tipo de Ocorrência

5.1.4. AnalysisAPI

Descrevemos anteriormente que uma das características de uma API Rest é ter comunicação a partir de requisições HTTP. Para isso, foi construído um módulo em Python que recebe essas

requisições e dá como resposta um objeto do tipo JSON[9] (*JavaScript Object Notation)*. O padrão de requisição utilizado no serviço sobre situação de endereço pode ser visto na Figura 10, já a Figura 11 apresenta um exemplo de resultado.

URL de Requisição

http://localhost/analysis_api/?**horario**=**NUMBER**&**data**=**DATE**&**endereco=ADDRESS**&**intervalo**=**60**

Parâmetros

endereco

String especificando o local de onde se deseja extrair as ocorrências.

horario

Horário da situação em que aconteceram as ocorrências do endereço especificado. Deve ser passada num formato *HH*:*mm*

data

Data da situação em que aconteceram as ocorrências do endereços especificado. Deve ser passada num formato dd/MM/yyyy

intervalo

Intervalo de tempo no qual o horário das ocorrências registradas no banco de Dados podem diferir do horário passado por parâmetro da URL.

Figura 10 - AnalysisAPI: padrão de requisição do serviço de situação de endereço

Exemplo de requisição

GET http://localhost/analysis_api/?**horario**=**8:00**&**data**=**20/03/2013**&**endereco=Rua+Conselheiro+Portela**&**intervalo**=**60**

Resultado

```
[
{
        'endereço' : Rua Conselheiro Portela,
        'CEP': 52020030,
        'data_hora': 2013-03-20 07:44:04,
        'categoria_ocorrencia' : 'negativa'
        'tipo_ocorrencia' : 'Semáforo defeituoso'
        'id_ocorrencia' : 13
        'id_tipo_ocorrencia' : 16
},
 {
        'endereço' : Rua Conselheiro Portela,
        'CEP': 52020030,
        'data_hora': 2013-03-20 08:30:09,
        'categoria_ocorrencia' : 'negativa'
        'tipo_ocorrencia' : 'Engarrafamento/Situação do Trânsito'
        'id_ocorrencia' : 20
        'id_tipo_ocorrencia' : 5
}
]
```

Figura 11 - AnalysisAPI: Resultado de requisição sobre situação de endereço.

[9] **JSON** <http://www.json.org/>

Já o padrão de requisição e resposta do serviço sobre mensagens é apresentado na Figura 12. A requisição é mais simples, pois só possui um parâmetro.

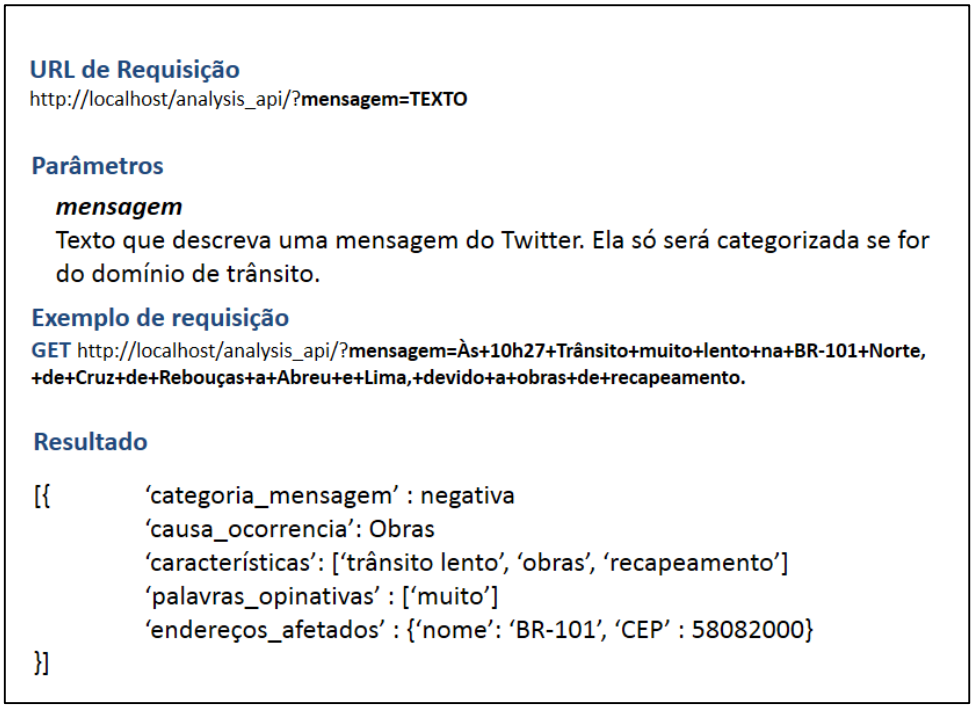


Figura 12 - AnalysisAPI: Serviço sobre Categorização de Mensagem

A partir de testes manuais, considerou-se a resposta no formato JSON como mais adequada, pois várias linguagens de programação e dispositivos conseguem interpretar objetos nesse formato, como desktops, browser, smartphones, entre outros.

5.2 Resultados obtidos

Como apresentado anteriormente, o UbibusAnalysis consiste, então, numa ferramenta de interpretação de mensagens de trânsito com Análise de Sentimentos. Após construída, desenvolveu-se um modelo de avaliação dos resultados.

O primeiro módulo de avaliação consistia na classificação das mensagens em relação à sua causa de ocorrência e sua categoria. Causa da ocorrência se refere ao atributo **tipo** das características – alagamento, acidente, situação de trânsito, entre outros (ver **Apêndice 7**). No segundo módulo de avaliação o objetivo foi verificar se a extração dos objetos (descoberta dos endereços onde acontecem as ocorrências) estava retornando os locais corretos.

Ainda há alguns desafios a serem enfrentados, e serão descritos no próximo capítulo, mas os resultados encontrados parecem satisfatórios, visto que não foi encontrada uma ferramenta com um foco igual ao do UbibusAnalysis. A avaliação da precisão/corretude da análise foi feita manualmente e os resultados são apresentados a seguir.

Na primeira avaliação, que leva em consideração a natureza das mensagens, 1280 (um mil duzentos e oitenta) mensagens foram avaliadas:

- Aproximadamente 88% (1128 mensagens) foram classificadas corretamente quanto à causa da ocorrência e sua categoria (positiva ou negativa).

- Aproximadamente 11% (151 mensagens) foram classificadas de maneira errada. Nesses casos o erro aconteceu quando as mensagens não tratavam exatamente de trânsito ou, ainda, quando eram perguntas ou ironias.

Com esses resultados, foi percebido que em trabalhos como os descritos no Capítulo 3, costuma-se pré-selecionar o conjunto de avaliação, removendo mensagens que não fazem parte do domínio. Se isso tivesse sido feito, os resultados poderiam ter subido a uma taxa de até 95% de acerto.

Na segunda avaliação, que leva em consideração a extração de objetos (endereços), foi avaliado um conjunto de 1135 (um mil cento e trinta e cinco) mensagens, que faziam parte já do primeiro conjunto:

- Aproximadamente 80% (910 mensagens) extraíram os endereços corretamente das mensagens.
- Os outros 20% (227 mensagens), não tiveram os endereços extraídos corretamente e se adequam a um outro nível de análise, onde há desafios a serem enfrentados.

5.3. Considerações

Depois de implementar o projeto do UbibusAnalysis e avaliar a classificação, consideramos os resultados bastante satisfatórios. A abordagem a partir de dicionários construídos manualmente se apresentou muito boa, pois o domínio de mensagens de trânsito é bastante específico.

Outra observação é que, visto que as mensagens a serem analisadas não são "pré-anotadas" como no trabalho de [AGARWAL et. Al. 2011], o sistema consegue lidar bem com mensagens que não fazem parte do domínio. Além disso, o algoritmo de similaridade entre *strings* se mostrou rápido e eficiente.

No próximo capítulo apresentamos uma conclusão e algumas perspectivas para a continuação deste trabalho.

6. Conclusão e Trabalhos Futuros

O objetivo deste trabalho foi atingido com sucesso. Foi criada uma ferramenta de análise que interpreta as mensagens de trânsito e provê informação útil para outros usuários e aplicações. Para isto, foram aprendidos conceitos novos como os de Análise de Sentimentos, uma área recente, até conceitos básicos de processamento de cadeias de caracteres, protocolos *web*, entre outros. Também foram aprendidas e estudadas novas tecnologias diferentes para a criação e uso de cada um dos componentes do UbibusAnalysis.

Contudo, apesar dos bons resultados, foram constatados alguns pontos de melhoria que não foram alcançados neste projeto e que ficam como sugestão para trabalhos futuros. Elas são listadas abaixo com uma breve justificativa.

- **Construir uma ferramenta de busca inteligente:**
  Atualmente, o TweetsCrawler faz buscas fixas em três perfis do *Twitter*. Uma sugestão seria adicionar algum tipo de aprendizado à busca, de maneira que ela – independentemente – buscasse por informações úteis e de trânsito em outros perfis. Além de incluir tratamento para outras redes sociais. O Facebook[10], por exemplo, já possui grupos que discutem especificamente de mensagens de trânsito.

- **Desenvolver um componente para construção do dicionário de características e expandir a extração delas:**
  O dicionário de características é fixo e foi construído manualmente. Embora ele leve em consideração algumas gírias e muitas palavras do vocabulário de trânsito, se houvesse uma ferramenta *inteligente* e capaz de *aprender*, a construção desse dicionário poderia ser dinâmica. Além disso, seria necessário um estudo aprofundado em processamento de linguagem natural para, talvez, cobrir casos de ironia e negação.

- **Expandir a etapa de Extração de Objetos**
  Na fase de avaliação, constatamos que a extração de objetos não extraía corretamente os endereços porque, muitas vezes, os *tweets* traziam nomes de bairros – o que confundia a análise –, bem como “apelidos” de ruas. Expandir a Base de Dados para conter apelidos de ruas pode ajudar nesse processo. Além disso, expandir o tratamento de extração para capturar dois endereços em uma mesma mensagem também pode aumentar a taxa de acerto na extração.

  Por outro lado, é necessário estudar e entender o contexto das fontes das mensagens, os problemas específicos de cada cidade. Sendo assim, quando um *tweet* mencionar um bairro, pode ser possível inferir quais os endereços que estão sendo afetados.

[10] Rede Social **Facebook** <https://www.facebook.com/facebook>.

Referências

[ABBASI ET AL. 2008] ABBASI, A., CHEN, H., SALEM, A. **SENTIMENT ANALYSIS IN MULTIPLE LANGUAGES: FEATURE SELECTION FOR OPINION CLASSIFICATION IN WEB FORUMS**. IN ACM TRANSACTIONS ON INFORMATION SYSTEMS (TOIS), V.26 N.3, P.1-34, JUNE 2008.

[AGARWAL ET. AL. 2011] AGARWAL, A., XIE B., VOVSHA I., RAMBOW O., PASSOUNNEAU R.. **SENTIMENT ANALYSIS OF TWITTER DATA**. IN PROCEEDINGS OF THE WORKSHOP ON LANGUAGES IN SOCIAL MEDIA, PAGES 30–38. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2011.

[AKKAYA & WIEBE 2009] AKKAYA, C., WIEBE J. AND MIHALCEA, R. **SUBJECTIVITY WORD SENSE DISAMBIGUATION**, EMNLP '09 PROCEEDINGS OF THE CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING: 190-199, 2009.

[AN ET AL. 2011] AN, S.-H., LEE, B.-H. AND SHIN, D.-R. **A SURVEY OF INTELLIGENT TRANSPORTATION SYSTEMS, COMMUNICATION SYSTEMS AND NETWORKS** (CICSYN), THIRD INTERNATIONAL CONFERENCE, PP.332- 337, 2011.

[BALDAUF 2007] BADAUF, M. **A SURVEY ON CONTEXT-AWARE SYSTEMS**, V-RESEARCH, INDUSTRIAL RESEARCH AND DEVELOPMENT, STADTSTRASSE 33, 6850 DORNBIRN, AUSTRIA 2007.

[BIFET & FRANK 2010] BIFET, ALBERT; FRANK, EIBE. **SENTIMENT KNOWLEDGE DISCOVERY IN TWITTER STREAMING DATA**. IN: DISCOVERY SCIENCE. SPRINGER BERLIN HEIDELBERG. P. 1-15, 2010.

[BIGONHA ET AL. 2010] BIGONHA, C. A. S., CARDOSO, T. N., MORO, M. M., ALMEIDA, V. A., & GONÇALVES, M. A. (2010). **DETECTING EVANGELISTS AND DETRACTORS ON TWITTER**. IN18TH BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB (PP. 107-114).

[BOYD & ELLISON 2011] BOYD, DANAH M.; ELLISON, NICOLE B.. **SOCIAL NETWORK SITES: DEFINITION, HISTORY AND SCHOLARSHIP**. JOURNAL OF COMPUTER MEDIATED COMMUNICATION, 2008. DISPONÍVEL EM: <HTTP://ONLINELIBRARY.WILEY.COM/DOI/10.1111/J.1083-6101.2007.00393.X/FULL>. ACESSO EM: 16/04/2013.

[BRÉZILLON 1999] BRÉZILLON, P. **CONTEXT IN ARTIFICIAL INTELLIGENCE: IA SURVEY OF THE LITERATURE**, COMPUTER & ARTIFICIAL INTELLIGENCE, V. 18, PP. 321-340, 1999.

[CUI ET AL. 2011] CUI, A., ZHANG, M. LIU, Y. AND MA, S. **EMOTION TOKENS: BRIDGING THE GAP AMONG MULTILINGUAL TWITTER SENTIMENT ANALYSIS**. IN: INFORMATION RETRIEVAL TECHNOLOGY. SPRINGER BERLIN HEIDELBERG. P. 238-249, 2011.

[FERNANDES, 2010] FERNANDES, F. **UM FRAMEWORK PARA ANÁLISE DE SENTIMENTO EM COMENTÁRIOS SOBRE PRODUTOS EM REDES SOCIAIS**. DISSERTAÇÃO (MESTRADO EM CIÊNCIA DA COMPUTAÇÃO) - CENTRO DE INFORMÁTICA/UFPE. RECIFE. 2010.

[GÓMEZ ET AL. 2009] GÓMEZ, A., DIAZ, G. AND BOUSETTA, K. **ITS FORECAST: GIS INTEGRATION WITH ACTIVE SENSORY SYSTEM**. IN: INFORMATION INFRASTRUCTURE SYMPOSIUM, GIJS'09 GLOBAL, 2009.

[LIMA ET AL. 2012] LIMA, V., MAGALHÃES, F., TITO, A. O., SANTOS, R., RISTAR, A., SANTOS, L., VIEIRA V. E SALGADO, A. C. **UBIBUSROUTE: UM SISTEMA DE IDENTIFICAÇÃO E SUGESTÃO DE ROTAS DE**

ÔNIBUS BASEADO EM INFORMAÇÕES DE REDES SOCIAIS. IN: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO 2012.

[LIU 2010.A] LIU, B. **SENTIMENT ANALYSIS AND SUBJECTIVITY. HANDBOOK OF NATURAL LANGUAGE PROCESSING**, SECOND EDITION, (EDITORS: N. INDURKHYA AND F. J. DAMERAU), 2010.

[LIU 2010.B] LIU, B. **SENTIMENT ANALYSIS: A MULTI-FACETED PROBLEM**. DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF ILLINOIS AT CHICAGO, 2010.

[PANDEY & IYER 2010] PANDEY, V., AND IYER, C. V. K. **SENTIMENT ANALYSIS OF MICROBLOGS**. SYSTEM. AVAILABLE AT: HTTP://WWW.STANFORD.EDU/CLASS/CS229/PROJ2009/PANDEYIYER.PDF, 2010.

[PANG & LEE 2008] PANG, B. AND LEE, L. **OPINION MINING AND SENTIMENT ANALYSIS**. FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL, 2008.

[PEREIRA ET AL. 2009] PEREIRA, D. ZILLER, J. **A IMOBILIDADE MÓVEL NOS ENGARRAFAMENTOS DE BELO HORIZONTE**. III SIMPÓSIO NACIONAL DA ABCIBER 2009.

[RUFINO 2009] RUFINO, AIRTIANE F**. TWITTER: A TRANSFORMAÇÃO NA COMUNICAÇÃO E NO ACESSO ÀS INFORMAÇÕES**. XI CONGRESSO DE CIÊNCIAS DA COMUNICAÇÃO NA REGIÃO NORDESTE. TERESINA (PI): INTERCOM, 2009.

[SANTANA 2007] SANTANA, CAMILA LIMA S. E. **REDES SOCIAIS NA INTERNET: POTENCIALIZANDO INTERAÇÕES SOCIAIS**. HIPERTEXTUS REVISTA DIGITAL (UFPE), V. I, P. 25-33, 2007.

[SILVA 2000] SILVA, DANYELA MORAES DA**. SISTEMAS INTELIGENTES NO TRANSPORTE PÚBLICO COLETIVO POR ÔNIBUS**. 2000. DISSERTAÇÃO (GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO) - UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. PORTO ALEGRE.

[SILVA FILHO 2011] SILVA FILHO, A. **REDES SOCIAIS NA ERA DA CONECTIVIDADE** ("THE GOOD, THE BAD AND THE UGLY"). REVISTA ESPAÇO ACADÊMICO, BRASIL, 10 DEZ. 2010. DISPONÍVEL EM: <HTTP://WWW.PERIODICOS.UEM.BR/OJS/INDEX.PHP/ESPACOACADEMICO/ARTICLE/VIEW/11864/6373>.

[SIQUEIRA, 2010] SIQUEIRA, H**. WHATMATTER: EXTRAÇÃO E VISUALIZAÇÃO DE CARACTERÍSTICAS EM OPINIÕES SOBRE SERVIÇOS**. DISSERTAÇÃO (MESTRADO EM CIÊNCIA DA COMPUTAÇAO) - CENTRO DE INFORMÁTICA/UFPE. RECIFE. 2010.

[SUSSMAN 2005] SUSSMAN, J. **PERSPECTIVES ON INTELLIGENT TRANSPORTATION SYSTEMS**. NEW YORK, USA: SPRINGER, 2005.

[TONG 2011] TONG R. **AN OPERATIONAL SYSTEM FOR DETECTING AND TRACKING OPINIONS IN ON-LINE DISCUSSION**. IN PROCEEDINGS OF THE ACM SIGIR WORKSHOP ON OPERATIONAL TEXT CLASSIFICATION, 1–6, 2001.

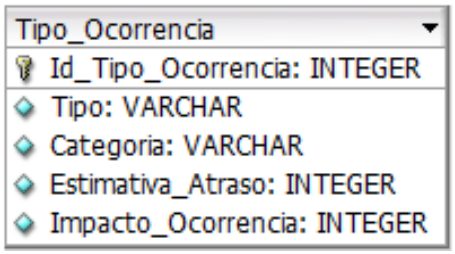
[VIEIRA ET AL. 2011] VIEIRA, V., CALDAS, L. R., SALGADO, A. C. **TOWARDS AN UBIQUITOUS AND CONTEXT SENSITIVE PUBLIC TRANSPORTATION SYSTEM**. IN: 4TH INTERNATIONAL CONFERENCE ON UBI-MEDIA COMPUTING (U-MEDIA 2011), 2011, SÃO PAULO, P.174 - 179

[WIEBE ET AL. 2005] WIEBE, T. WILSON, AND C. CARDIE. 2005. **ANNOTATING EXPRESSIONS OF OPINIONS AND EMOTIONS IN LANGUAGE**. LANGUAGE RESOURCES AND EVALUATION (FORMERLY COMPUTERS AND THE HUMANITIES), 164–210.

[ZIMMERMANN ET AL. 2007] ZIMMERMANN, A., LORENZ, A., OPPERMANN, R. **AN OPERATIONAL DEFINITION OF CONTEXT**, IN: PROC. OF THE 6TH INTERNATIONAL AND INTERDISCIPLINARY CONFERENCE ON MODELING AND USING CONTEXT, PP. 558-571, ROSKILDE, DENMARK, 2007.

Apêndice 1 – Modelagem Conceitual do Ubibus Completa.

Apêndice 2 – Padronização e detalhes da tabela Tipo Ocorrência



Para atender às diversas necessidades do projeto Ubibus [VIEIRA et al. 2011] e do projeto UbibusAnalysis, foram estabelecidas algumas regras de povoamento da tabela **Tipo Ocorrência**.

Regras:

1. O atributo **categoria** da tabela **Tipo_Ocorrencia**, tem somente 2 valores possíveis: Positivo ou Negativo. Ocorrências de categoria Positivo, tem impacto sempre com valor 3 ou 4. Já as de categoria Negativo, tem impacto sempre de 1 ou 2.

2. O atributo **impacto_ocorrencia**, será a pontuação de uma ocorrência, com valores numéricos de *1* a *4*, representando um determinado nível do trânsito. **Quanto menor a pontuação do impacto, pior a situação do** trânsito em termos de congestionamentos ou retenções.
3. O atributo **estimativa_atraso** não foi utilizado pelo UbibusAnalysis.
4. O atributo **tipo** pode assumir os seguintes valores:
    - Alagamento
    - Engarrafamento/Situação do Trânsito
    - Buraco
    - Acidente
    - Semáforo defeituoso
    - Protesto
    - Passeata
    - Bloqueio
    - Obras

Sendo assim, o povoamento feito e utilizado no UbibusAnalysis se encontra na Tabela **Tipos de Ocorrência.**

id_tipo_ocorrencia (*integer* NOT NULL)	tipo (*text*)	categoria (*text*)	estimativa_atraso (*integer*)	impacto_ocorrencia (*integer*)
1	'Alagamento'	'Negativo'	0	1
2	'Alagamento'	'Negativo'	0	2
3	'Alagamento'	'Positivo'	0	3
4	'Engarrafamento/ Situação do Trânsito'	'Negativo'	0	1
5	'Engarrafamento/ Situação do Trânsito'	'Negativo'	0	2
6	'Engarrafamento/ Situação do Trânsito'	'Positivo'	0	3
7	'Engarrafamento/ Situação do Trânsito'	'Positivo'	0	4
8	'Buraco'	'Negativo'	0	1
9	'Buraco'	'Negativo'	0	2
10	'Buraco'	'Positivo'	0	3
11	'Buraco'	'Positivo'	0	4
12	'Acidente'	'Negativo'	0	1
13	'Acidente'	'Negativo'	0	2
14	'Acidente'	'Positivo'	0	3
15	'Semáforo defeituoso'	'Negativo'	0	1
16	'Semáforo defeituoso'	'Negativo'	0	2
17	'Semáforo defeituoso'	'Positivo'	0	3
18	'Protesto'	'Negativo'	0	1
19	'Protesto'	'Negativo'	0	2
20	'Protesto'	'Positivo'	0	3
21	'Passeata'	'Negativo'	0	1
22	'Passeata'	'Negativo'	0	2
23	'Passeata'	'Positivo'	0	3
24	'Bloqueio'	'Negativo'	0	1
25	'Bloqueio'	'Negativo'	0	2
26	'Bloqueio'	'Positivo'	0	3
27	'Obras'	'Negativo'	0	1
28	'Obras'	'Negativo'	0	2
29	'Obras'	'Positivo'	0	3

Tabela de Tipos de Ocorrência

Apêndice 3 – Stopwords do tipo Pontuação e Caracteres especiais

A lista de stopwords do tipo Pontuação é:

.
:
;
?
!
(
)
/
\\
-
&
*
,
%

Apêndice 4 – Stopwords de pré-processamento

a	com	isso	proximo
as	esta	este	que
o	transito	aquele	partir
os	ponto	aquela	partida
um	frente	aquilo	passar
uma	tras	logo	ir
algum	distante	area	via
alguns	ponto	envolver	motivo
na	momento	envolvendo	inicio
no	causa	cttu	esse
em	devido	desde	essa
para	ambos	vindo	por
pra	sentido	indo	pelo
pois	sentidos	local	chegar
foto	imediacoes	proibicao	entrar
fotos	so	proibido	sair
porque	via	proibir	direcao
pq	cruzamento	nas	fumaca
acho	outro	nos	quando
acha	outros	ja	

Apêndice 5 – Stopwords secundárias

carro	melhor	bem	eletrica	motociclista
moto	pior	operacao	estabelecida	ficar
carros	fluxo	prossegue	poste	ferido
motos	estacao	prosseguir	calcada	deixar
onibus	metro	sofrer	pedestre	faixa
cachorro	detalhe	sofrerao	bateria	continuar
cao	carnaval	confira	altura	por
policia	festa	autopasseio	segue	conta
caminho	palco	transito	seguir	proibicao
carreta	sao joao	RT	desbloqueio	sem
hospital	tudo	entre	andamento	alimentacao
pista	nada	gerar	acomodando	fugir
estrada	todo	gera	acomodar	cair
caminhao	mais	carregar	auto	passar
carreta	frente	descarregar	cruzamento	e
evite	atras	descarregando	semaforo	situacao
informacao	cruzamento	morador	sinal	ao
informacoes	perigo	moradores	placa	apos
saida	completo	fechar	colegio	deslocar
sentido	completar	fecham	escola	deslocando
motociclista	mineral	tanto	fila	problema
motoqueiro	cachorro	quem	dupla	diversos
Ciclista	vc	quando	tripla	recapeamento
grande	voce	fora	depois	celta
veiculo	nunca	dentro	antes	prata
veiculos	acesso	sincronia	durante	iniciam
chuva	ciclista	completamente	banco	comecar
tem	bicicleta	bairros	cidade/suburbio	comecam
de	devido	vizinhos	cidade	bom
do	ambos	bairro	suburbio	penitenciario
da	comecando	vizinho	suburbio/cidade	bandido
uma	terminando	patrulha	padaria	sent.
uma	funciona	policia	final	quebrado
uns	funcionando	delegacia	comeco	inteiro
umas	quebrado	retorno	fim	capotar
ocorrendo	amassa	ainda	iniciar	capotagem
ocorre	cai	sua	chegar	deslize
acontecendo	fere	suas	sair	deslizamento
acontece	foge	seu	vitima	lombada
costamento	parar	motorista	condutor	

Apêndice 6 – Dicionário de Palavras Opinativas

Palavras Opinativas	*Categoria*
muito	*positivo*
muita	***positivo***
muitos	***positivo***
pouco	***negativo***
poucas	***negativo***
pouca	***negativo***
pouco	***negativo***
bastante	***positivo***
nada	***negativo***
nenhum	***negativo***
sem	***negativo***
grande	***positivo***
super	***positivo***
completamente	***positivo***
totalmente	***positivo***

Apêndice 7 – Dicionário de Características

Tipo	Alagamento
Prefixos	['alag']
Derivados	['alagamento', 'alagado', 'pista molhada', 'chuva', 'derramamento', 'encosta', 'neblina']
Categoria	Negativo
Impacto_Característica	2
Impacto_Mínimo	1
Impacto_Máximo	3

Tipo	Engarrafamento/Situação do Transito
Prefixos	['parad', 'interdit']
Derivados	['parado', 'parada', 'triste', 'pessimo', 'pessimo', 'caos', 'caotico']
Categoria	Negativo
Impacto_Característica	1
Impacto_Mínimo	1
Impacto_Máximo	2

Tipo	Engarrafamento/Situação do Transito
Prefixos	['engarraf', 'congest', 'complic', 'intens', 'lent']
Derivados	['engarrafado', 'engarrafada', 'complicado', 'complicada', 'pesado', 'pesada', 'enrolado', 'enrolada', 'intenso', 'intensa', 'lento', 'lenta', 'devagar', 'ruim', 'atrapalhado', 'engarrafamento', 'moderado', 'retido', 'retencao', 'retencoes', 'fluxo', 'insuportavel']
Categoria	Negativo
Impacto_Característica	2
Impacto_Mínimo	1
Impacto_Máximo	3

Tipo	Engarrafamento/Situação do Transito
Prefixos	['fluind', 'tranqu', 'livr', 'rapid']
Derivados	['fluindo', 'tranquilo', 'tranquila', 'facil', 'rapido', 'rapida', 'livre', 'normal', 'leve', 'bom', 'flui bem', 'flui', 'normalmente']
Categoria	Positivo
Impacto_Característica	3
Impacto_Mínimo	3
Impacto_Máximo	4

Tipo	Buraco
Prefixos	['burac', 'buraqu']
Derivados	['buraco', 'buracos', 'buraqueira', 'cratera', 'asfalto irregular', 'problema calcamento']
Categoria	Negativo
Impacto_Característica	2
Impacto_Mínimo	1
Impacto_Máximo	3

Tipo	Acidente
Prefixos	['acident', 'atropelam', 'tomb']
Derivados	['acidente', 'batida', 'choque', 'carro quebrado','engavetamento', 'tombamento', 'carreta tombada', 'atropelamento', 'atropelado', 'colidir', 'colisao', 'bateu', 'tombar']
Categoria	Negativo
Impacto_Característica	2
Impacto_Mínimo	1
Impacto_Máximo	3

Tipo	Semáforo defeituoso
Prefixos	['semaforo', 'sinal']
Derivados	['semaforo defeituoso', 'semaforo defeito', 'semaforo quebrado', 'sinal quebrado', 'sinal defeito', 'sinal']
Categoria	Negativo
Impacto_Característica	2
Impacto_Mínimo	1
Impacto_Máximo	3

Tipo	Protesto
Prefixos	['protesto']
Derivados	['protesto']
Categoria	Negativo
Impacto_Característica	2
Impacto_Mínimo	1
Impacto_Máximo	3

Tipo	Passeata
Prefixos	['passeat']
Derivados	['passeata']
Categoria	Negativo
Impacto_Característica	2
Impacto_Mínimo	1
Impacto_Máximo	3

Tipo	Bloqueio
Prefixos	['bloq', 'interd', 'interromp']
Derivados	['interditado', 'bloqueado', 'bloqueio', 'bloqueamento', 'interdicao', 'interromper', 'interrompida']
Categoria	Negativo
Impacto_Característica	2
Impacto_Mínimo	1
Impacto_Máximo	3

Tipo	Obras
Prefixos	['obra']
Derivados	['obras', 'obra', 'construcao', ‘recapeamento’]
Categoria	Negativo
Impacto_Característica	2
Impacto_Mínimo	1
Impacto_Máximo	3