UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

ESCOLA DE INFORMÁTICA APLICADA

Análise de Sentimento dos Jogos Olímpicos Rio 2016 no Twitter

Carolina Yorio Heinze Tozzi

Orientadoras

Kate Revoredo

Flávia Santoro

RIO DE JANEIRO, RJ – BRASIL

DEZEMBRO DE 2016

Análise de Sentimento do Jogos Olímpicos Rio 2016

no Twitter

Carolina Yorio Heinze Tozzi

Projeto de Graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada por:

_____

Kate Revoredo, D.Sc. (UNIRIO)

_____

Flávia Maria Santoro, D.Sc. (UNIRIO)

_____

Fernanda Araujo Baião, D.Sc. (UNIRIO)

_____

Vânia Maria Felix Dias, D.Sc. (UNIRIO)

RIO DE JANEIRO, RJ – BRASIL.

DEZEMBRO DE 2016

Agradecimentos

Agradeço, primeiramente, a minha mãe Simone, minha tia Sueli e meu padrasto Paulo por me apoiarem quando decidi largar a carreira em química para começar uma nova em Sistemas de Informação. Sem o apoio deles eu nunca teria começado o curso.

Queria agradecer às professoras Kate e Flávia pela orientação neste trabalho, todas as dicas, opiniões, correções e paciência ao longo do desenvolvimento do projeto.

Queria agradecer também aos amigos que estiveram presente durante as duas graduações que eu fiz pelos momentos especiais, trabalhos de final de semana e toda a troca de conhecimento de vida e acadêmico.

RESUMO

A análise de dados para descoberta de conhecimento recebeu grande atenção nas últimas décadas. As redes sociais são excelentes fontes de dados por estarem disponíveis a qualquer momento, em qualquer lugar e pelo fato do usuário poder opinar sobre qualquer assunto. Isso torna as redes sociais excelentes plataformas para análise de sentimento e de opinião. Por conta da enorme quantidade de dados gerados pela utilização de redes sociais, a análise manual de sentimentos presentes nesses dados é inviável. A mineração de dados é uma solução para analisar dados de redes sociais. Ela faz parte da descoberta de conhecimento em dados estruturados e não estruturados. Os dados provenientes das redes sociais são não estruturados e precisam de um pré-processamento para que o conhecimento contido neles seja extraído. Uma das plataformas mais utilizadas para mineração de dados em redes sociais é o Twitter. Ele fornece uma API que permite a captura de tweets públicos para a análise de conteúdo. O presente trabalho apresenta uma análise de sentimento durante os Jogos Olímpicos Rio 2016. Saber o que os usuários do Twitter pensam a respeito dos jogos e o sentimento desses usuários, sejam brasileiros ou não, foi a motivação da análise proposta. A análise foi feita em três períodos diferentes (antes, durante e depois dos jogos) e para dois idiomas (inglês e português). A cada dia durante os três períodos, 500 tweets com a hashtag oficial dos jogos (#Rio2016) foram capturados e uma captura de tweets em tempo real também foi realizada diariamente. Fazer a análise manual, lendo cada um dos tweets de cada período seria extremamente trabalhoso, demorado e passível de muitos erros por parte do analista. A mineração de dados minimiza o tempo de análise, a quantidade de trabalho necessário e os possíveis erros proveniente de cansaço.

Palavras-chave: Análise de sentimento, Mineração de dados, Olimpíadas, Twitter

ABSTRACT

The data analysis for knowledge discovery got a lot of attention in the past few decades. Social Networks are excellent data sources because they are available at all times, all places and for the fact that the user can give an opinion on any subject. Those factors makes social networks great platforms for sentiment and opinion analysis. Due to the huge amount of data generated by the usage of social networks, a manual sentiment analysis of the data would not be feasible. Data mining is one of the solution to analyze the data. Data mining is a part of the knowledge discovery in structured and unstructured data. Social network data are unstructured and require a pre-processing to extract the knowledge within it. Twitter is one of the most used platforms in data mining. Twitter has an API that allows the capture of the public tweets to analyze its contents. This project presents a sentiment analysis of the Olympic Games in Rio 2016. Knowing what Twitter users think about the games and those users' feelings, Brazilian or not, was the main motivation of the proposed analysis. The analysis was made in three different periods of time (before, during and after) and for two languages (English and Portuguese). On each day of the three periods, 500 tweets with the Games official hashtag (#Rio2016) were captured and a real time capture was also made on each day. Do a manual analysis reading each tweet of each period would be a extremely hard work, would take a long time and a lot of mistakes could be made by the analyst. Data mining minimizes the analysis time, the amount of work and the errors, especially due to tiredness.

Keywords: Sentiment analysis, Data mining, Olympics, Twitter

Índice

Índice de Tabelas

Índice de Figuras

1 Introdução

1.1 Motivação

Nos últimos anos, o volume, a variedade e a velocidade dos dados aumentaram muito, além de serem provenientes de diversas fontes. Entre outros fatores, o aumento no volume de geração de dados foi impulsionado pela onipresença da web, o armazenamento de memória mais barato e a consciência de que o dado é a peça chave para a descoberta de novos conhecimentos [1].

Inteligência de negócios e análise (BI&A) e o campo de análise de big data se tornaram incrivelmente importantes tanto na comunidade acadêmica quanto nas empresas nas duas últimas décadas. Uma pesquisa feita pelo IBM Tech Trends Report identificou BI&A como uma das 4 tendências tecnológicas na década de 2010 [2].

As redes sociais são uma frente dessa revolução, apresentando vários desafios associados com big data [1]. Hoje em dia, todos podem publicar opiniões, visões, ideias e interesses sobre os tópicos que quiserem a qualquer hora de qualquer lugar do mundo. As redes sociais podem ser usadas para discutir o mercado financeiro, vendas e possíveis surtos de doenças [3, 4, 5, 6].

A análise de sentimento é uma das áreas de pesquisa mais ativa em processamento de linguagem natural, mesmo esse processamento sendo apenas a etapa inicial de tratamento necessária para realizar a análise. Esse aumento na importância da análise de sentimento se deu graças ao crescimento das redes sociais [7]. Utilizar análise de dados pode ajudar, por exemplo, a entender como o público em geral se sentiu em relação a um determinado evento, pode ajudar a encontrar pontos fortes e fracos da organização do evento que não foram perceptíveis à organização e com isso apoiar as pessoas envolvidas a tomarem melhores decisões no futuro.

Este projeto de graduação foi motivado por quatro tópicos: análise de dados, modelagem de processo, processo de software e a grande quantidade de dados disponíveis que poderiam ser utilizados. A Figura 1 mostra um diagrama que resume as motivações para o desenvolvimento do projeto. A ideia geral do projeto é apresentar e discutir o desenvolvimento de um projeto de análise de grande volume de dados em redes sociais, através de um processo de execução bem definido, e que pode ser replicado em outros contextos.

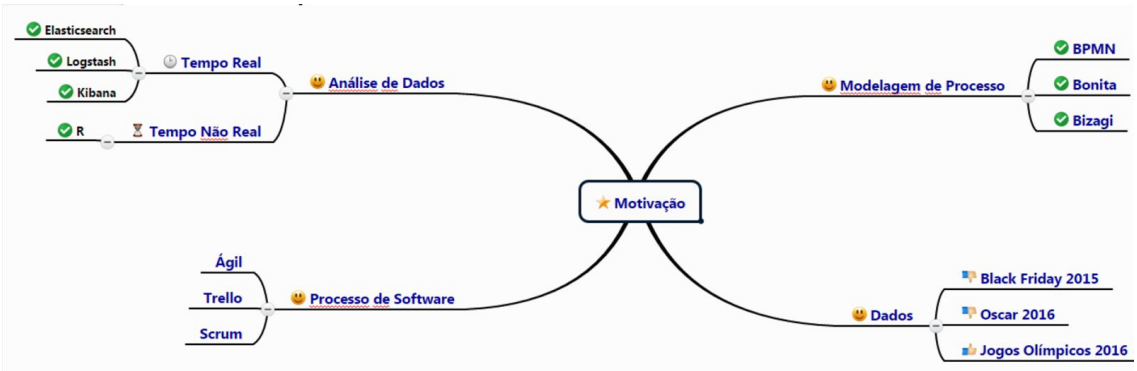

Figura 1- Motivações para o desenvolvimento do projeto

Na perspectiva da análise de dados, aprender ferramentas que fizessem coleta de dados de redes sociais em tempo real e não real e conseguir gerar conhecimento através da análise desses dados foram a motivação do trabalho. A geração do conhecimento envolve também a visualização dos dados, ou seja, como os dados são apresentados ao usuário. E desta forma, os Jogos Olímpicos foram então escolhidos como evento para coleta de dados e análise.

Na perspectiva de modelagem de processo, o aprofundamento na notação Business Process Modeling Notation (BPMN) [8] e o contato com ferramentas de modelagem utilizadas pelo mercado foi o foco do trabalho. Um projeto de desenvolvimento necessita de uma modelagem dos processos envolvidos, de uma análise do quão próximo da realidade o processo está. Neste projeto, o modelo construído representa todas as etapas que foram necessárias para realizar a análise de sentimento através de dados do Twitter.

A metodologia escolhida foi a metodologia Ágil utilizando os conceitos de Scrum. Esta metodologia foi escolhida por trabalhar com entregas menores, ou seja, existem entregas semanais no projeto, seja de um método novo necessário ou uma mudança nas ferramentas utilizadas, e mais importante, permite correções mais

rápidas, mudança de curso mais rápidas. Se algo que estava planejado não deu certo ou não teve os resultados esperados, na semana seguinte era possível pensar em uma solução para o problema ao invés de só saber do problema em uma fase mais avançada do projeto.

1.2 Objetivos

Este trabalho tem como objetivo principal o estudo e aplicação de técnicas para capturar tweets sobre as olimpíadas e fazer análise de sentimento desses tweets em três momentos diferentes - antes, durante e depois das olimpíadas - e utilizando a hashtag #Rio2016. A intenção é saber se o sentimento das pessoas mudou ao longo do evento, se as reclamações e elogios mudaram, se problemas reportados pelos moradores da cidade e visitantes foram resolvidos ao longo do evento ou não.

Os objetivos secundários são:

- Capturar tweets em tempo real e em tempo não real utilizando as ferramentas e técnicas necessárias para cada um dos tipos de captura;
- Criar um modelo de processo para a análise realizada, desde a configuração dos softwares até a descoberta do conhecimento resultante da análise.

O modelo de processo gerado ajudou a manter a consistência dos passos necessários, principalmente no pré-processamento dos tweets e na análise deles, durante todo o tempo do desenvolvimento do projeto para evitar que alguns desses passos fossem esquecidos ou que alguma das etapas definidas no início se mostrassem desnecessárias no futuro. A criação do modelo do processo também pode ajudar na realização de análises futuras, documentando de maneira mais visual tudo que é necessário para o método de análise proposto.

Foi estabelecida a utilização de uma metodologia Ágil para a realização do projeto, isto é, planejar as atividades do projeto para que existam entregas contínuas de tarefas e de uma maneira que mudanças ao longo do projeto não prejudiquem o planejamento inicial.

A metodologia Ágil escolhida para ser usada no projeto foi Scrum. No Scrum os projetos são baseados em ciclos chamados iterações que podem ser mensais ou semanais. Essas iterações são blocos de tempo em que atividades são realizadas.

As funcionalidades necessárias ao projeto são listadas em uma lista de tarefas. No começo de cada iteração uma reunião de planejamento (iteração planning) é realizada e as atividades que serão feitas naquela iteração são definidas.

Ao final de cada iteração uma reunião é feita para que os membros do time discutam o que deu certo, errado e o que pode ser melhorado para a próxima iteração. Essa é a retrospectiva (iteração retrospective) [9]. A Figura 2 ilustra o processo da metodologia Scrum.

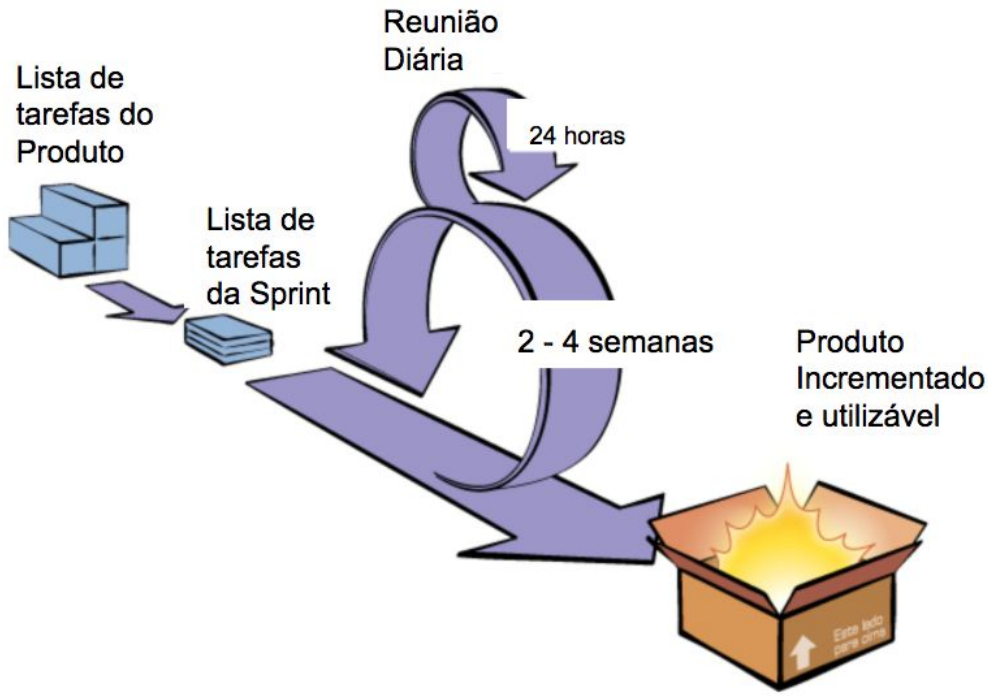

Figura 2- Processo da metodologia Scrum

O projeto foi dividido em 5 grandes áreas que foram colocadas como lista de tarefas do projeto. Cada grande área foi dividida em atividades e essas atividades tinham 3 estágios: “a fazer”, “fazendo” e “feito”. A realização das atividades foram divididas em iterações de 1 semana ou seja, as atividades que passavam do “a fazer” para o “fazendo” seriam terminadas em 1 semana e então passariam para o “feito”.

1.3 Organização do texto

O presente trabalho está estruturado em capítulos e, além desta introdução.

A revisão bibliográfica aborda cada um dos assuntos estudados neste trabalho e das ferramentas utilizadas para a realização dele.

O capítulo seguinte descreve todas as ferramentas que foram utilizadas durante o desenvolvimento do projeto.

A metodologia vai descrever todas as etapas da análise, desde a configuração das ferramentas até os códigos que analisaram os tweets frente aos dicionários.

Resultados e discussão vai analisar os resultados obtidos no trabalho. Abordará os resultados obtidos nas capturas de tweets, criação de dicionários e análises.

As conclusões vão pontuar as contribuições do trabalho e os possíveis estudos futuros que poderiam ser continuados.

2 Revisão Bibliográfica

2.1 Mineração de dados coletados em redes sociais

O processo de descoberta de conhecimento em banco de dados é o processo responsável por, a partir de uma base de dados, encontrar algum conhecimento. Ele é dividido em 7 passos [10].

- Limpeza dos dados - remover ruídos e dados inconsistentes;
- Integração dos dados - combinar várias fontes de dados;
- Seleção de dados - extrair do banco os dados relevantes à análise;
- Transformação dos dados - transformar os dados no formato apropriado para o tipo de análise;
- Mineração dos dados - identificar padrões que representem conhecimento baseado em métricas;
- Apresentação do conhecimento - mostrar para o usuário as visualizações e o conhecimento obtido com as técnicas de mineração .

A mineração de dados pode ser vista como a evolução natural da tecnologia da informação. Ela tem o potencial de transformar grandes quantidades de dados em conhecimento. A mineração de dados é tratada por algumas pessoas como descoberta de conhecimento de dados e por outros como parte dessa descoberta [10].

Os dois objetivos principais da mineração de dados são a descrição e a previsão. A descrição busca padrões existentes nos dados analisados enquanto a previsão usa dados existentes para prever dados faltantes ou dados futuros[11].

Um exemplo de algoritmo de mineração usado para classificação é o ID3. Ele constrói árvores de decisão baseado em um conjunto de dados de treinamento de amostras já classificadas. Para cada nó da árvore, o atributo que melhor divide o conjunto em duas categorias com maior ganho de informação é escolhido. Essa etapa se repete para as partições menores da árvore[12].

Um outro exemplo de algoritmo de mineração é o k-means. O algoritmo realiza iterações sempre em 2 passos:

1. Dividir os dados - ligar cada dado ao centróide mais próximo dele;
2. Recalcular as médias (ou dado utilizado para definir a distância até o centróide) - mover os centróides para a posição médias dos dados ligados a eles;

Esses dois passos se repetem até que nenhum dado mude de centróide depois de as médias serem recalculadas[12].

2.2 Mineração de texto

Descoberta de conhecimento pode ser definida como “identificar, receber informações relevantes e poder processá-las e agregá-las ao conhecimento prévio do usuário”[13]. O conjunto de técnicas utilizadas para organizar e descobrir conhecimento em bases textuais é denominada mineração de texto[14].

O campo da mineração de texto cresceu muito nos últimos anos por conta dos avanços tecnológicos de hardware e software para redes sociais possibilitando a criação de grande repositórios de vários tipos diferentes de dados[15]. Enquanto o dado estruturado é normalmente gerenciado por banco de dados, os textos usam mecanismos de busca pela falta de estrutura[15].

Mineração de texto é a parte do processo de descoberta de conhecimento que utiliza técnicas de análise e extração de dados a partir de textos, frases ou palavras[13].

O processo de mineração de texto envolve as etapas de seleção de documentos, definição do tipo de abordagem dos dados (semântica ou estatística), preparação dos dados, indexação e normalização, cálculo da relevância dos termos, seleção dos termos e pós-processamento[13].

A abordagem semântica é baseada na funcionalidade dos termos encontrados nos textos e a estatística na frequência dos termos. O trabalho utilizou as duas abordagens, com foco maior na estatística do que na semântica.

A preparação dos dados é responsável pela seleção inicial do núcleo que melhor expressa o conteúdo dos textos.

A indexação e normalização geram um índice de termos utilizados nos documentos. Após a indexação, processos para identificar termos (simples e

compostos), remoção de *stopwords* (palavras muito comuns para serem relevantes) e stemming (normalização morfológica) são aplicados.

A Figura 3 mostra um fluxograma das etapas normalização dos termos.

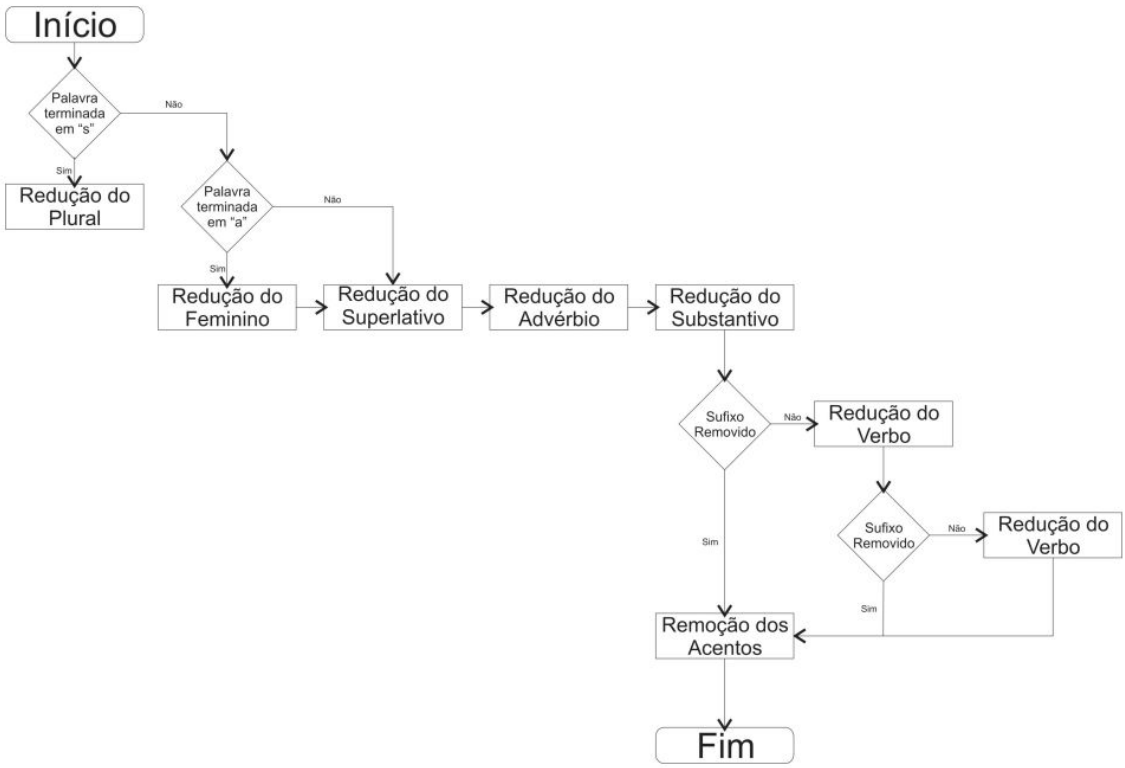

Figura 3- Stemming para texto em português[13]

As técnicas de descoberta de conhecimento em texto podem ser divididas em supervisionadas e não supervisionadas.

Nas técnicas supervisionadas, o algoritmo utilizado recebe os exemplos a serem analisados e os resultados esperados e ele tenta se adaptar para apresentar o resultado esperado para cada exemplo.

As técnicas não supervisionadas apenas dados, não existem modelos ou exemplos para aprendizado do algoritmo, ele fica responsável por encontrar relacionamentos entre os dados[16].

2.3 Redes Sociais

Nos últimos anos, as redes sociais online se tornaram uma enorme fonte de informações. Isso é devido, principalmente, ao fato de as publicações serem em tempo

real e as pessoas expressarem opiniões, discutirem assuntos pertinentes e notícias recentes[17]. Alguns conceitos de análise de redes como ator, laços relacionais, dupla, trio, grupo, subgrupo, relacionamento e rede são fundamentais para entender as redes sociais[18].

Ator pode ser uma pessoa, corporação ou mesmo uma unidade social. Esses atores são ligados a outros atores por laços relacionais que podem ser tipos e abrangência diferentes. Dupla é a ligação entre dois atores e muitas análises usam a dupla como foco de estudo. O trio é a ligação entre três atores e pode ser classificada em transitiva ou balanceada. O subgrupo é um conjunto de atores e laços sem importar as quantidades de cada um deles. Um grupo pode ser um conjunto de atores em que os laços serão avaliados, ou seja, é o conjunto que tem o foco da análise. A relação é um conjunto de laços do mesmo tipo que está presente no grupo[18].

Uma rede social consiste em um grupo finito de atores e as relações definidas entre eles[18]. Os estudos convencionais em redes sociais não são focados em interações online, eles precedem historicamente a popularização dos computadores e da internet[19].

A disponibilidade de grandes quantidade de dados online deu um novo rumo às pesquisas estatísticas em redes sociais[19]. Sites como Twitter[1], Facebook[2], Tumblr[3], recebem milhões de postagens diárias sobre os mais diversos assuntos. O formato livre das postagens e o fácil acesso aos microblogs fazem os usuários da internet mudarem dos blogs tradicionais e emails para os microblogs como principal meio de comunicação[20].

O Twitter é uma fonte rica de dados, ter os dados à disposição do público e uma API bem documentada, faz dele um ótimo ponto de partida para mineração de dados em redes sociais. Os dados do Twitter são interessantes por acontecerem na velocidade do pensamento e estarem disponíveis quase instantaneamente[21].

2.4 Análise de sentimento

Informação textual pode ser dividida em dois grandes grupos: fatos e opiniões. Fatos são expressões objetivas sobre eventos e propriedades. Opiniões são, normalmente, expressões subjetivas que descrevem o sentimento das pessoas, apreciações e sentimentos sobre entidades, eventos e seu pertences[22].

[1] https://twitter.com/
[2] https://www.facebook.com/
[3] https://www.tumblr.com/

O estudo de opiniões, sentimentos e emoções expressas em texto é conhecido como análise de sentimento[16].

Acredita-se que no ano de 2001 a atenção se voltou aos problemas da pesquisa e às oportunidades que a análise de sentimento e a mineração de opinião poderiam trazer. O desenvolvimento do aprendizado de máquinas no processamento de linguagem natural, a disponibilidade de banco de dados para treinamento de algoritmos e os desafios intelectuais envolvendo aplicações comerciais e inteligente que a área traz motivaram trouxeram interesse para a área[23].

Hoje em dia, análise de sentimento é parte integral do monitoramento de redes sociais. A abordagem usual é dizer “se um conteúdo tem mais palavras positivas que negativas, ele é positivo; se tiver mais palavras negativas que positivas, ele é negativo”. Não é incorreto falar isso, mas não é tão simples[24].

Análise de sentimento é um problema de processamento de linguagem natural (PLN). Está em todos os aspectos da PLN, resolução de correferência, lidar com negação e desambiguação de palavras, o que aumenta a dificuldade da análise pois esses problemas ainda não foram solucionados[25].

Em [26] foi feita a análise de sentimento em tempo real de um jogo de futebol da copa das confederações de 2013. Observou-se que a quantidade de tweets aumentou nos eventos importantes do jogo como a defesa de um pênalti ou o gol da vitória. A polaridade dos comentários também seguiu mudando de acordo com os eventos do jogo. Um dos eventos que teve uma polaridade negativa acentuada foi uma falta desnecessária cometida por um zagueiro[26].

Uma análise de sentimento das eleições presidenciais dos Estados Unidos. O estudo revelou que os eventos das campanhas aumentavam o número de tweets postados. Um dos discursos feitos pelo, candidato na época, Barak Obama obteve o maior número de tweets em 5 meses de análise. Esse discurso aumentou de 3 a 4 vezes a polaridade de tweets tanto negativo quanto positivo se comparado a um dia normal[27].

Em outro exemplo deste tipo de pesquisas, Thelwall et al (2009), analisou a rede social MySpace, avaliando a presença de comentário positivos e negativos e o gênero do autor dos comentários. Dois terços dos comentários analisados mostraram sentimento positivo contra um terço de comentários negativos e o gênero feminino teve maior tendência em receber e fazer mais comentários positivos que o gênero masculino[28].

2.5 API do Twitter

O twitter disponibiliza duas APIs para captura de dados da rede social: REST API - captura tweets sem ser em tempo real - e Streaming API - captura tweets em tempo real. As duas têm integração com linguagens como python, R e Java e integração com softwares como Elasticsearch.

Essas APIs permitem que os tweets sejam capturados em tempo real ou com no máximo 7 dias desde a data de postagem. Ela permite ainda que parâmetros como geolocalização, idioma, tipo de resultado (tempo real, mais popular ou misturado), número de tweets a serem retornados, definição de intervalo de tempo para a captura dos tweets (considerando um tempo máximo de 7 dias antes do momento da busca), entre outros[29].

Para a utilização das APIs, é necessário ter uma conta no Twitter. Acessando o site www.dev.twitter.com é possível criar uma aplicação, gerar os 4 códigos de acesso necessários para que a API funcione. O *consumer key secret* e o *access token secret* devem permanecer secretos pois são responsáveis pela autenticação de segurança da aplicação[29].

Considerando todo o cenário descrito neste capítulo uma ideia de propor uma forma de analisar dados do Twitter da maneira mai automática possível surgiu. Para tanto, uma série de ferramentas foram estudadas e escolhidas para serem utilizadas neste trabalho.

3 Ferramentas de análise de dados

3.1 Linguagem R

R é uma linguagem e um software gratuito para cálculos estatísticos e criação de gráficos. Ela é similar à linguagem S, pode ser considerada uma implementação diferente do S[30]. Um dos pontos fortes da linguagem R é a facilidade de produzir plots de alta qualidade, incluindo símbolos matemáticos.

A linguagem R já tem pacotes especiais para determinados tipos de análise de dados, o CRAN Task Views é um bom guia. Entre as análises encontram-se[31]:

- Aprendizagem de máquina estatístico;
- Análise de cluster e modelos probabilísticos;
- Estatística multivariada;
- Análise de dados espaciais.

Um dos pacotes do aplicativo R é o twitteR responsável por fazer a conexão entre o software e a API do Twitter para captura de dados. Este pacote tem funções que podem ser usadas para buscar informações públicas de qualquer usuário - como os tweets publicados - e informações apenas para o usuário da conta - se determinado usuário segue você e para gerenciamento das mensagens diretas, por exemplo[32].

A autenticação da aplicação é feita utilizando-se o pacote ROAuth. Esse pacote permite que os usuários se façam autenticação nos servidores escolhidos[33].

A Figura 4 mostra um exemplo de código usado neste projeto para capturar tweets pelo R usando o pacote twitteR.

```
> library("twitteR")
  library("ROAuth")
  setup_twitter_oauth(consumer key,consumer secret,access_token= SEU
  ACCESS TOKEN,access_secret=SEU ACCESS TOKEN SECRET)
  search.string <- "#Rio2016"
  no.of.tweets <- 10
  tweets <- searchTwitter(search.string, n=no.of.tweets, lang="en")
```

Figura 4- Código de captura de tweets pelo R

As duas primeiras linhas importam os pacotes necessários para a busca de tweets: twitteR e ROAuth. A função *setup_twitter_oauth* é a responsável por fazer a autenticação do usuário e da aplicação na API do Twitter. A função *searchTwitter* é a responsável por buscar e retornar os tweets com os parâmetro definidos nela. No exemplo acima, ela retornaria 10 tweets que tivessem a palavra "#Rio2016" e salvaria no objeto tweets.

3.2 Elasticsearch, Logstash, Kibana

O Elasticsearch é uma ferramenta para busca e análise de dados. É um servidor de dados que suporta vários formatos de linguagens de buscas. Ele é escalável, confiável e fácil de usar[34].

O Logstash recebe logs de diversas fontes, sendo o Twitter uma delas, e indexa essas informações no Elasticsearch. O Kibana é uma plataforma de visualização e criação de gráficos[34].

O Elasticsearch tem conexão com a API do twitter, é possível coletar tweets apenas com ele, mas ele é um sistema tela preta. O Logstash faz a conexão entre o Elasticsearch e o Kibana, permitindo assim que os tweets capturados pelo Elasticsearch sejam visualizados e tratados pela interface gráfica do Kibana.

A interface gráfica do Kibana permite que os tweets coletados sejam vistos em tempo real. É possível plotar gráficos que são atualizados também em tempo real.

A Figura 5 exemplifica um arquivo de configuração do Logstash.

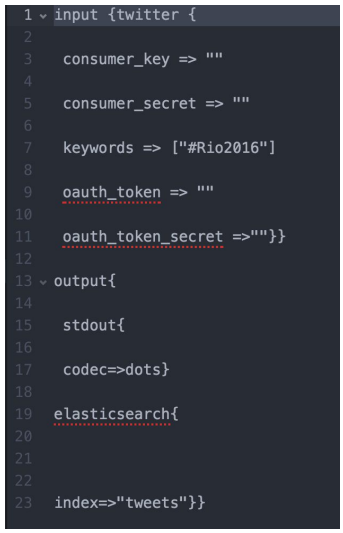

Figura 5- Arquivo de configuração do Logstash

As duas primeiras informações do arquivo são os códigos da aplicação do Twitter consumer key e consumer secret. *Keywords* é um array que contém todas as palavras que serão buscadas nos tweets, pois não precisa ser uma só. As duas informações seguintes são relativas ao token também da aplicação do Twitter. O output define o tipo de configuração que será usado para compactação e descompactação dos dados que serão enviados ao Kibana e o index é a identificação desta configuração.

A interface gráfica do Kibana utilizará o index do arquivo .conf para identificar as informações que devem ser mostradas. O index deverá ser escolhido na aba de settings e os tweets aparecerão na aba de discover. Em seguida, uma visualização para cada um dos termos do index pode ser criada e elas podem ser visualizadas ao mesmo tempo no dashboard. As Figuras 6 a 9 mostram as interfaces do Kibana.

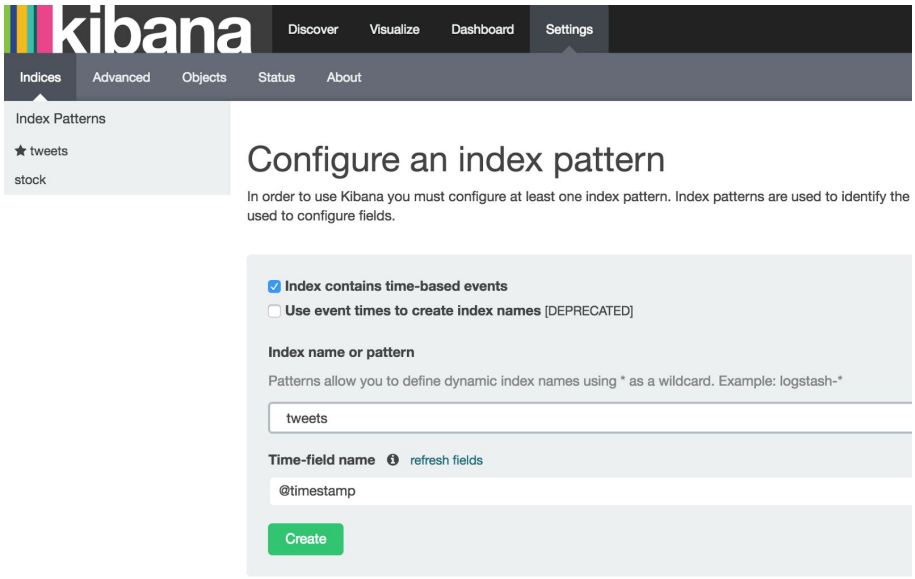

Figura 6- Página de configuração do Kibana

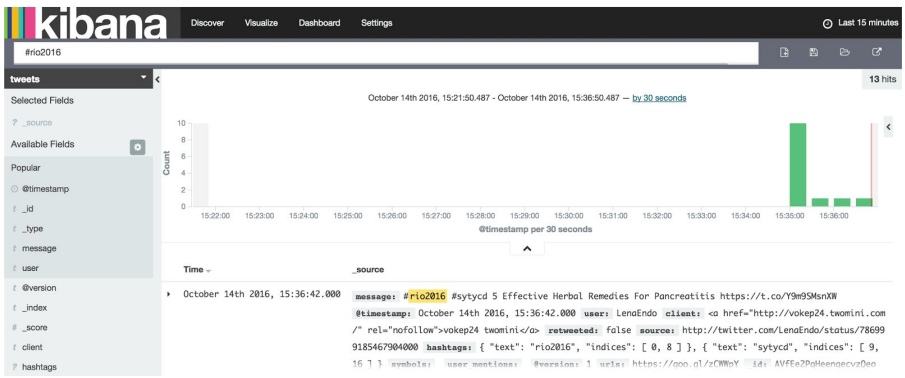

Figura 7- Página de descoberta do Kibana

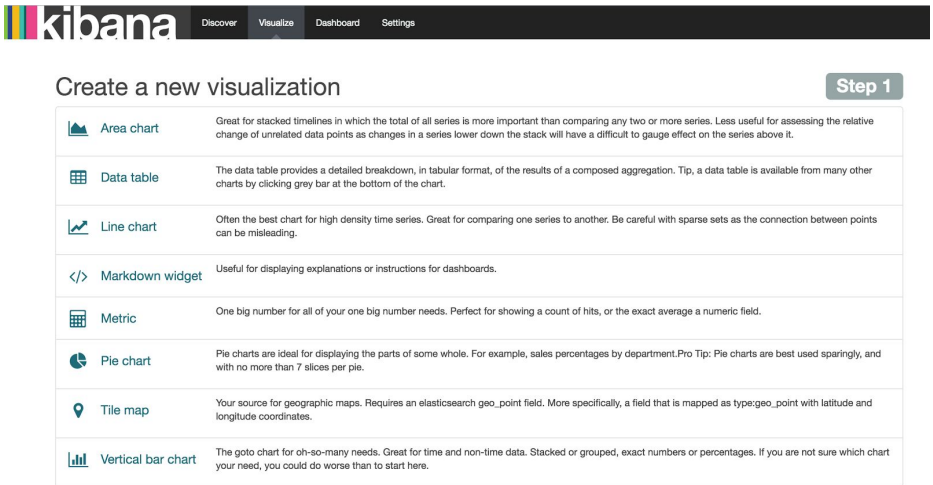

Figura 8- Página de visualização do Kibana

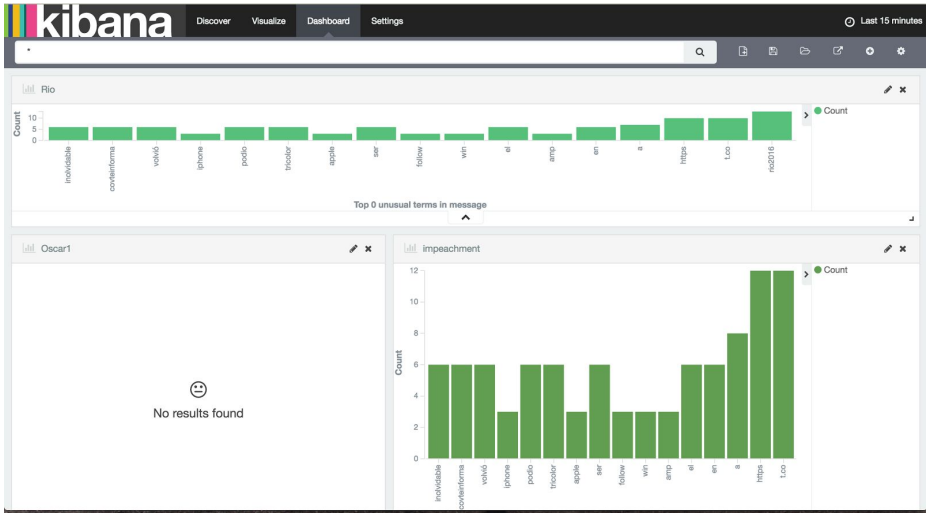

Figura 9- Página de dashboard do Kibana

3.3 Watson Analytics

O Watson Analytics é um serviço de análise e predição de dados utilizando o computador cognitivo da IBM, Watson. É um serviço na nuvem, ou seja, pode ser acessado de qualquer computador.

O serviço tem um teste grátis que pode ser usado por qualquer pessoa, basta cadastrar a conta com um endereço de email válido. Para criar a conta é necessário informar o email, uma senha de 8 dígitos, primeiro e último nomes, nome de uma empresa (pode ser a universidade) e um número de telefone. Um código será enviado ao seu email e ele é necessário para conseguir completar o primeiro acesso a sua conta.

A versão trial do serviço tem algumas limitações como tamanho máximo de arquivos e de armazenamento, só funciona para conta individual, não para times e dura 30 dias. As funcionalidades da versão trial são um pouco diferentes em relação às versões paga e para IBMistas.

A Figura 10 mostra a homepage do Watson Analytics.

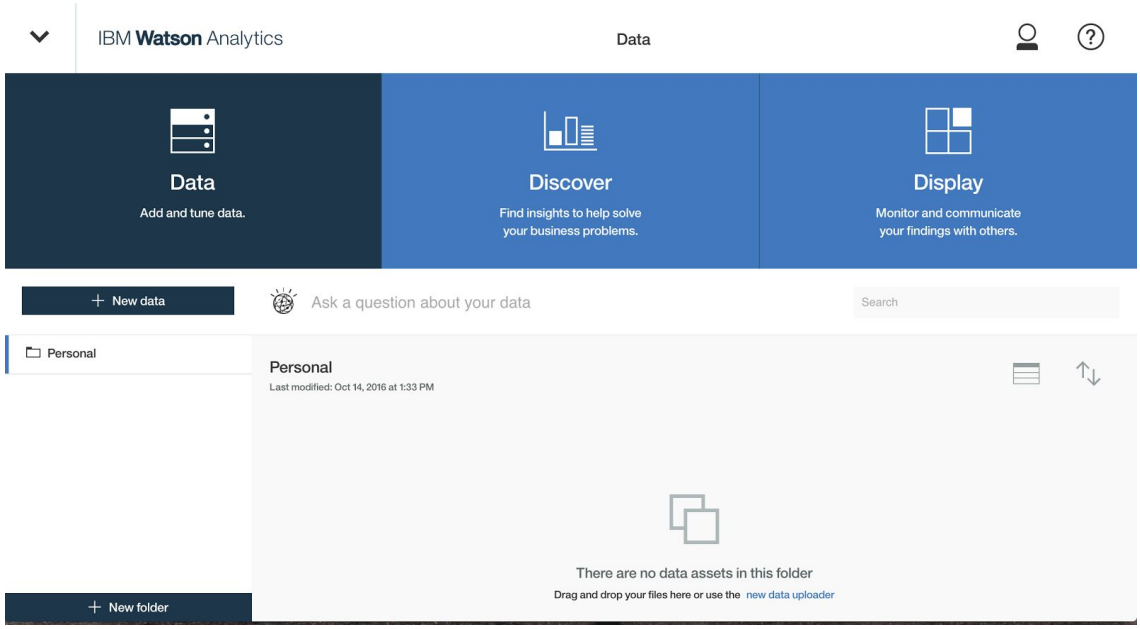

Figura 10- Home do Watson Analytics

O primeiro passo é criar um novo conjunto de dados. Ele pode ser importado de plataformas como DropBox, Cognos e OneDrive. Conexões com banco de dados também podem ser utilizadas nos serviços pagos ou para IBMistas. E existe também a opção de importar um arquivo local ou usar os dados de exemplo já existentes no Watson Analytics.

É possível utilizar uma das visualizações sugeridas pelo Watson Analytics ou criar a que for mais adequada para o que seus dados descrevem. As visualizações dos discoveries podem ser mudadas dentro de um mesmo discover ou podem ser criados vários discoveries para o mesmo set de dados.

3.4 Alchemy

O Alchemy é uma API que oferece diversas opções de análise de texto. A análise pode ser feita a partir de HTML, arquivos de texto ou uma URL pública. O Alchemy pode ser usado através de um demo online (https://alchemy-language-demo.mybluemix.net/), no Bluemix (plataforma IBM de desenvolvimento) ou o código pode ser baixado do GitHub em node.js (https://github.com/watson-developer-cloud/alchemylanguage-nodejs)[35].

A versão gratuita do Alchemy tem algumas limitações como o número de requisições feita pela API no Bluemix e pelo código tem um limite diário e o demo online só funciona para textos em inglês.

A função de extração de keywords do Alchemy retorna todas as palavras relevantes do texto ou URL analisada e o sentimento associado a ela. O resultado das análises pode ser extraído em diversos formatos para o código do GitHub ou pelo Bluemix. O demo online retorna apenas o JSON dos resultados, a tabela que é gerada com os resultados também pode ser copiada e colada em um arquivo.

A Figura 11 mostra a interface do Alchemy online.

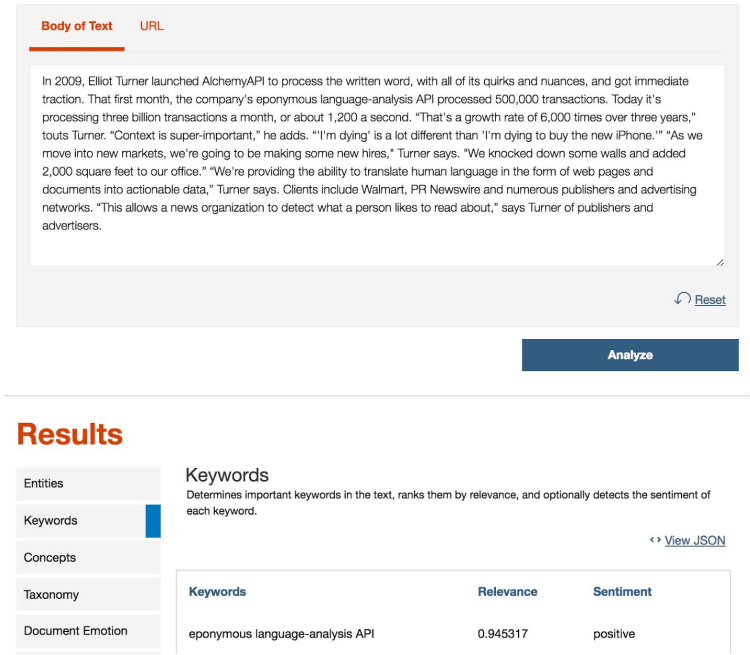

Figura 11- Interface do Demo online do Alchemy

Uma arquitetura com as ferramentas apresentadas foi definida no desenvolvimento deste projeto. Ela tem a finalidade de criar dicionários, um conjunto de palavras que expressam um determinado sentimento. A criação dos dicionários em inglês foi feita no Alchemy pois ele já fornecia o sentimento atrelado a cada palavra, facilitando a escolha de quais palavras estariam em cada dicionário. Como o demo online do Alchemy só suporta análises em inglês, a combinação do pacote ELK junto com o Watson Analytics foi utilizada na criação dos dicionários em português. O ELK para a captura dos tweets em tempo real e o Watson para analisar a frequência

dos termos extraídos pelo ELK. A Figura 12 apresenta um esquema da arquitetura das ferramentas utilizadas no projeto.

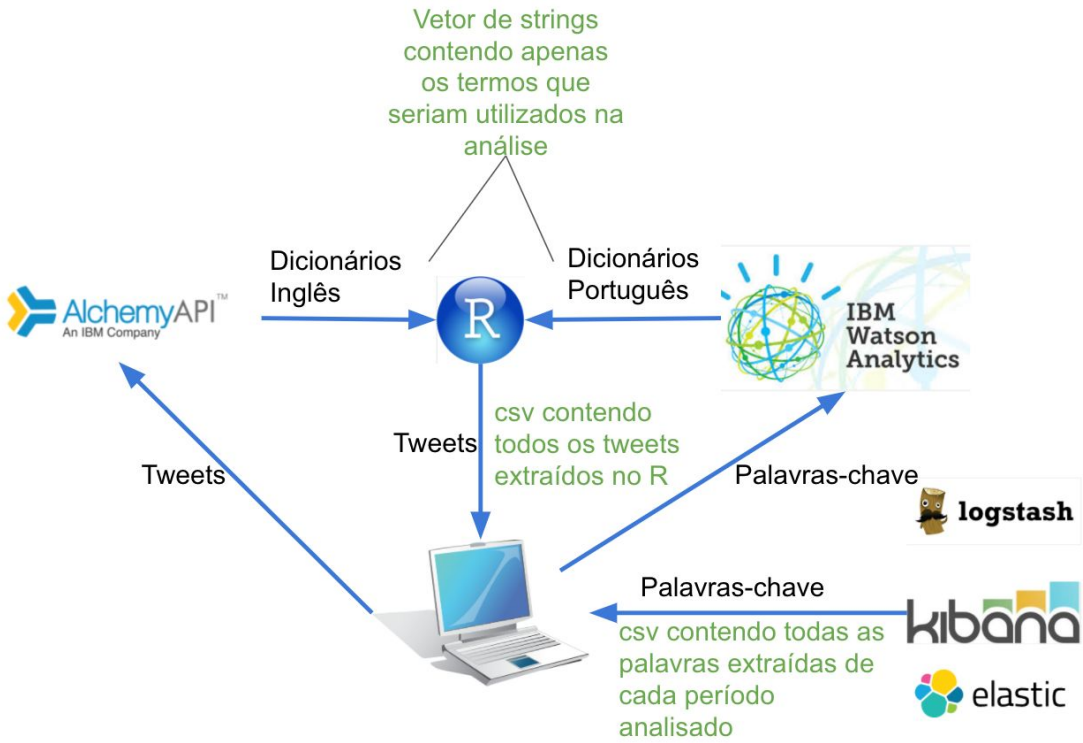

Figura 12 - Arquitetura das ferramentas utilizada no projeto

A escolha das ferramentas utilizadas no projeto foram feitas com base no artigo da IBM [26] e pesquisas realizadas antes e durante o projeto. O esquema mostrado acima é a infraestrutura final utilizada no projeto. O R sendo a ferramenta central de análise, o pacote ELK junto com o Watson Analytics gerando as palavras-chave em portugês e o Alchemy gerando palavras chave em inglês.

4 Metodologia

Para a execução do trabalho em questão, por ter atividades diárias e atividades semanais, utilizou-se a metodologia Ágil para execução das tarefas. Reuniões ou trocas de email, caso o encontro não fosse possível, com as orientadoras do projeto eram realizadas para acompanhamento. Nessas reuniões todas as atividades que haviam sido realizadas na semana anterior eram revisadas, caso precisassem ser modificadas (como algumas foram) elas entravam para a iteração da semana seguinte. Durante o desenvolvimento, algumas ferramentas, que não tinham sido cogitadas anteriormente, ganharam foco durante as reuniões e os testes delas entraram para o "a fazer" do projeto.

A Figura 13 é o quadro que mostra como estava o andamento das atividades antes do começo da coleta de tweets dia 5/7.

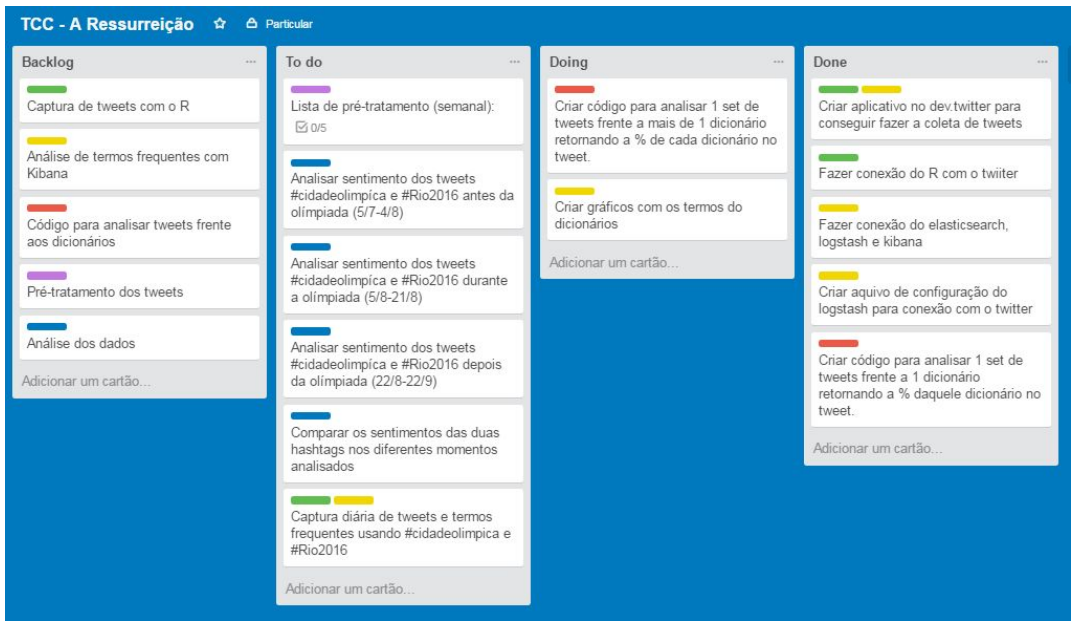

Figura 13- Quadro de atividades antes da coleta de tweets

Atividades de configuração das conexões, tanto do R quanto do Kibana já haviam sido finalizadas. A modificação do código de análise do conjunto de tweets frente aos dicionários estava sendo feita. E a definição de quais gráficos poderiam ser usados para representar melhor os termos mais frequentes estava sendo feita. As atividades que envolviam coleta, pré-tratamento e análise dos dados estavam na lista a ser realizada.

Ao longo das semanas, algumas tarefas foram concluídas, algumas foram descartadas e algumas começaram a ser realizadas. Na semana entre os dias 05/08 e 12/08, o quadro de tarefas estava de acordo com a Figura 14.

Figura 14- Quadro de atividades durante a coleta de tweets

Após o término das coletas e com as análises quase terminadas, na semana do dia 07/10 e 14/10, o quadro encontrava-se de acordo com a Figura 15.

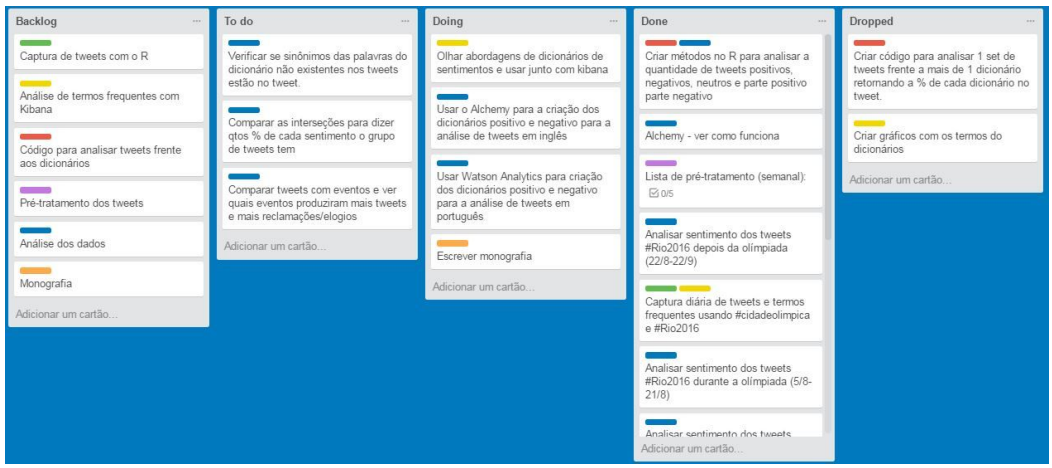

Figura 15- Quadro de atividades após a coleta de tweets

As seções seguintes descrevem o detalhamento dos fluxos de atividades do projeto através de modelos de processos em BPMN.

4.1 Configuração das ferramentas

A primeira parte do processo realizado no projeto foi a configuração das ferramentas. A criação de uma conta de desenvolvedor na API do Twitter é necessária para que a captura, em tempo real ou não, funcione. Uma aplicação foi criada nessa conta de desenvolvedor e as credenciais necessárias para a captura dos tweets são geradas nela.

Para o R foi necessário baixar e instalar 3 pacotes: twitteR, ROAuth. O pacote twitteR é responsável pela conexão entre o R e a API de desenvolvedor criada anteriormente. O pacote ROAuth é responsável por fazer o setup das credenciais da API no R, sem esse setup não é possível fazer a captura.

Para a configuração da combinação do Elasticsearch, Logstash e Kibana é preciso baixar o pacote ELK. Para que o Logstash faça a conexão entre o Elasticsearch e o Kibana é preciso criar um arquivo .conf. Este arquivo contém informações com as credenciais da aplicação do Twitter, a palavra buscada (pode ser mais de uma palavra) e uma variável responsável por conectar os tweets deste determinado arquivo à busca que aparece na visualização do Kibana.

A Figura 16 ilustra o processo de configuração das ferramentas.

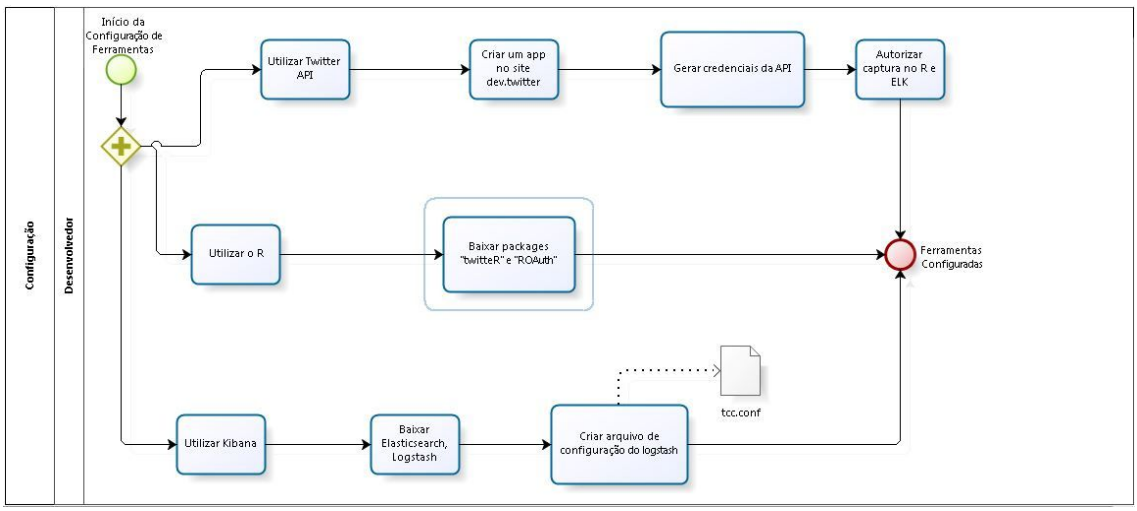

Figura 16- Processo de configuração das ferramentas

4.2 Captura dos Tweets

Para a captura com o pacote ELK é necessário iniciar os três programas no terminal, após a inicialização deles, é possível conectar a URL do Kibana no navegador. Na interface do Kibana, uma discovery foi criada para a busca específica contida no arquivo .conf. Uma visualização dessa discovery foi criada. Na visualização é possível criar os gráficos desejados, sendo que cada gráfico é uma visualização diferente. Esses gráficos podem ser histograma, área, barras entre outros como mostra a Figura 8. Um dashboard também foi criado e nele era possível ver todas as visualizações criadas. Os termos relevantes de cada dia de busca foram salvos em arquivos .csv. que eram armazenados localmente no computador.

A Figura 17 ilustra o processo de captura dos tweets em tempo real e não real.

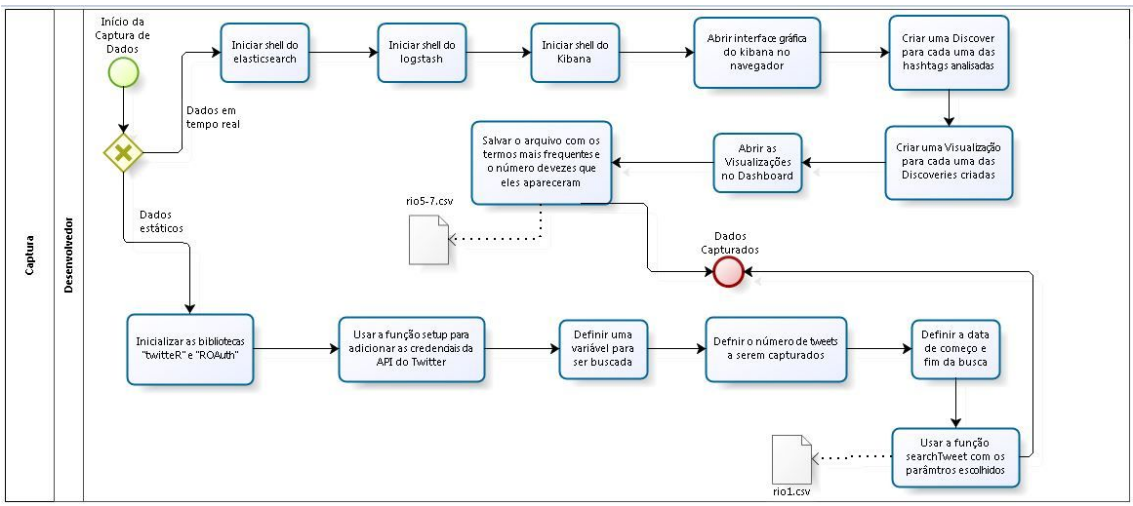

Figura 17- Processo de captura dos tweets.

Para a captura com o R, o primeiro passo é carregar os pacotes twitteR e ROAuth. Para a captura dos tweets é utilizado o método *searchTwitter().* Neste método podem ser definidos todos os parâmetros de busca que existem na API. Foram utilizados como parâmetros a palavra-chave “#Rio2016”, número de tweets foi setado em 500, o idioma inglês e português e o dia de coleta também foi utilizado como parâmetro. Os tweets coletados foram salvos em um arquivo .csv que era armazenado localmente no computador.

Neste trabalho, os metadados dos tweets coletados foram descartados uma vez que apenas o conteúdo das mensagens seria analisado. Para isso, criou-se um dataframe com os tweets coletados e esse dataframe original foi cortado para que

apenas a coluna “text” fosse mantida e o tamanho do objeto a ser trabalhado fosse menor.

4.3 Pré-Processamento

Para o pré-processamento, o pacote tm precisa ser instalado. Todos os tratamentos são feitos no objeto *corpus.* Esse objeto é uma coleção de documentos, cada tweet é um documento.

O dataframe cortado foi transformado em um *Vector Corpus.* O objeto *corpus* precisou ser colocado em UTF-8 para que erros fossem evitados. A Figura 18 mostra a análise de um dos objetos corpus criados no projeto. O corpus continha 500 tweets. É possível saber quantos caracteres cada tweet tem com o método “inspect” e qual o texto de cada um dos tweets com o método “writeLines”.

```
> corpus_rio1
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 500
>
> inspect(corpus_rio1[1:3])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 3

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 117

[[2]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 117

[[3]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 115

> writeLines(as.character(corpus_rio1[[2]]))
RT IndignadoRJ Faltam 30 dias   Olimpíadas

No primeiro dia  patrulha  cidade  carro  força nacional   alvo  tiros
```

Figura 18- Objeto corpus

Transformações para retirar pontuação, retirar as chamadas *“stop words”,* que são palavras muito comuns para serem relevantes, fazer o “steming” de palavras, que é transformá-las em seu radical (por exemplo, a palavra peixinho vira peixe), e deixar em texto corrido foram aplicadas ao objeto *corpus*.

Após as transformações, o *corpus* foi transformado em uma matriz de DocumentosVsTermos (DTM). A Figura 19 mostra um exemplo de DTM utilizado no projeto. A DTM tem 500 tweets, 1507 palavras diferentes e os dois primeiros tweets da lista contém a palavra “alvo”.

Essa matriz faz a correlação entre todos os termos e todos os documentos, caso um um documento contenha um termo, a posição da matriz correspondente ao documento e termo recebe o número 1, caso não contenha, recebe 0. A DTM foi utilizada para a separação dos tweets por sentimento contido nele.

```
> dtm_rio1
<<DocumentTermMatrix (documents: 500, terms: 1507)>>
Non-/sparse entries: 5693/747807
Sparsity           : 99%
Maximal term length: 25
Weighting          : term frequency (tf)
> inspect(dtm_rio1[1:3,50:59])
<<DocumentTermMatrix (documents: 3, terms: 10)>>
Non-/sparse entries: 2/28
Sparsity           : 93%
Maximal term length: 11
Weighting          : term frequency (tf)

    Terms
Docs alfinetadas alguma ali aliou alma aluno alvo amanhã amigo amigos
   1           0      0   0     0    0     0    1      0     0      0
   2           0      0   0     0    0     0    1      0     0      0
   3           0      0   0     0    0     0    0      0     0      0
```

Figura 19- Objeto document term matrix

A Figura 20 apresenta as atividades relacionadas ao pré-processamento dos tweets.

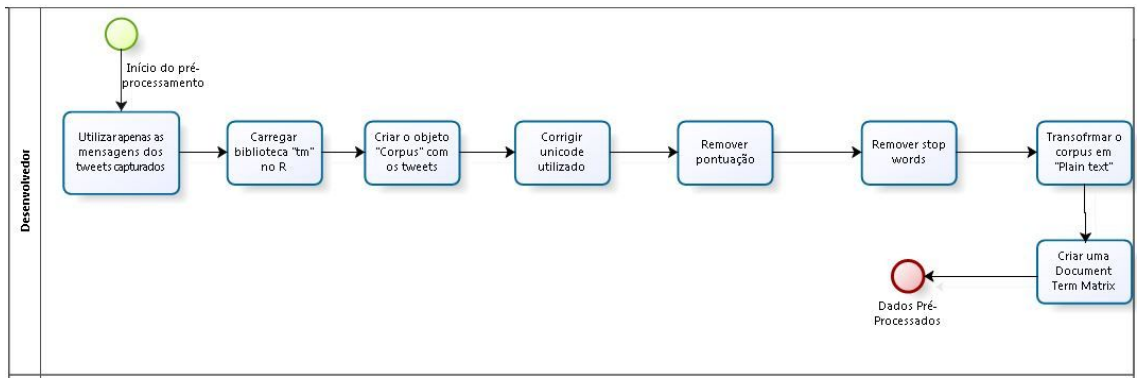

Figura 20- Pré-processamento dos tweets

4.4 Criação dos dicionários

Para a criação dos dicionários em português foram utilizados os tweets coletados em tempo real. Os tweets foram analisados todos os dias no período de (5/7 a 22/09) e dicionários positivo e negativo foram criados.

Em cada dia de análise, um arquivo .csv foi gerado pelo Kibana. Estes arquivos continham todos os termos relevantes de todos os tweets capturados e a quantidade de ocorrências de cada termo.

A Figura 21 ilustra o processo de criação dos dicionários.

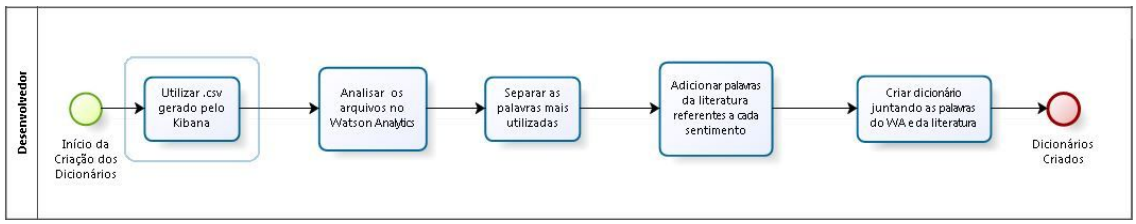

Figura 21- Processo de criação dos dicionários

Para a criação dos dicionários em inglês, o mesmo processo utilizado para o dicionário em português seria utilizado, mas optou-se pela utilização da ferramenta Alchemy para a extração dos termos relevantes para que o processo de criação dos dicionários fosse o menos manual possível. Os tweets capturados pelo R, separados por período, foram analisados pelo Alchemy. Um dicionário para sentimentos positivos foi criado com os termos relevantes que foram classificados como positivos. Um dicionário para sentimentos negativo foi criado com termos relevantes que foram classificados como negativos.

Os dicionários foram criados separadamente para cada momento de análise - antes, durante e depois das olimpíadas para a hashtag #Rio2016.

4.5 Análise frente aos dicionários

Após a criação dos dicionários, um método criado em R foi utilizado para fazer a comparação do texto dos tweets com as palavras dos dicionários. Este método conta quantas palavras tem um tweet e guarda essa quantidade. Em seguida, para cada palavra do dicionário encontrada no tweet um contador é aumentado em 1.

Feita a análise de todas as palavras, uma porcentagem que representa quanto de sentimento do dicionário (positivo e negativo) estava presente naquele tweet foi

calculada, dividindo-se a quantidade de palavras do dicionário presentes no tweet pelo total de palavras naquele tweet.

Após a comparação dos tweets frente aos dois dicionários, um segundo método foi utilizado para contar quantos tweets estavam presentes em cada uma das seguintes categorias:

- Positivo - um tweet que continha palavras apenas do dicionário positivo;
- Negativo - um tweet que continha palavras apenas do dicionário negativo;
- Neutro - um tweet que não continha palavras dos dicionários;
- Positivo e negativo - um tweet que continha palavras dos dois dicionários positivo;

A Figura 22 ilustra o processo utilizado na análise dos tweets.

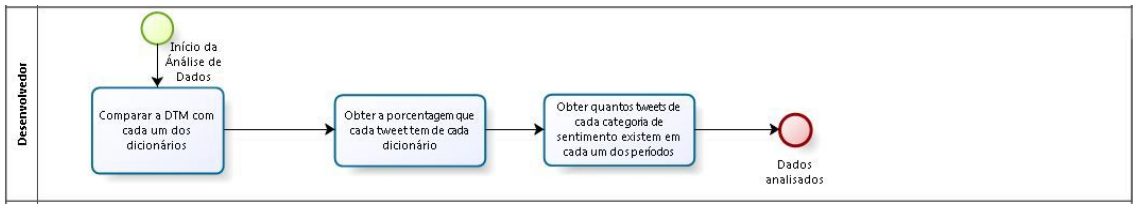

Figura 22- Processo de análise dos tweets

5 Resultados e Discussão

5.1 Metodologia Ágil

A utilização da metodologia Ágil na execução do projeto foi essencial para que todas as etapas fossem cumpridas. Atrasos na realização de algumas tarefas ocorreram, mas foi possível reorganizar a lista de “a fazer” do projeto e fazer com que no final as tarefas essenciais ao projeto tivessem sido completadas.

Algumas tarefas foram descartadas ao longo da execução do projeto para que outras com maior prioridade fossem realizadas, por exemplo, a criação de um código no R que comparasse os tweets a mais de um dicionário de uma só vez e a criação de gráficos com os termos dos dicionários foram descartadas e tarefas como entender como o Alchemy funcionava e se ele poderia servir aos propósitos do projeto foram adicionadas ao “a fazer” com maior prioridade.

A maioria das iterações foi realizado com sucesso e todas as tarefas propostas foram realizadas, sendo tarefas diárias - captura de tweets - ou semanais - pré-processamento dos tweets da semana.

As iterações que envolviam análise não foram tão bem sucedidas. Alguns atrasaram por conta de tarefas levando mais tempo do que o esperado. A utilização do Alchemy, principalmente, levou mais tempo do que o esperado pois a ferramenta era relativamente instável. O Alchemy aceitava em torno de 500 tweets por vez para análise. Muitas das vezes dava erro ao analisar os tweets ou, depois de analisar com sucesso, continuava mostrando os termos relevantes da análise anterior.

A iteração em que a tarefa de aprender a utilizar o Watson Analytics foi realizada também foi atrasada. A ferramenta não tem uma boa experiência de usuário, portanto para conseguir entender o que era preciso fazer para obter os resultados esperados levou mais tempo que o previsto.

5.2 Processos

O processo de criação dos dicionários foi alterado ao longo da execução do projeto. A princípio seria utilizado o mesmo processo utilizado para português: termos relevantes do Kibana e Watson Analytics.

Como o Alchemy se mostrou muito promissor na extração de termos relevantes, optou-se por utilizá-lo para a criação dos termos em inglês. Essa mudança pode ser considerada uma melhoria no processo já que o Alchemy extrai os termos relevantes já associados aos sentimentos positivo e negativo.

A Figura 23 ilustra como ficou o processo de criação dos dicionários com a utilização do Alchemy.

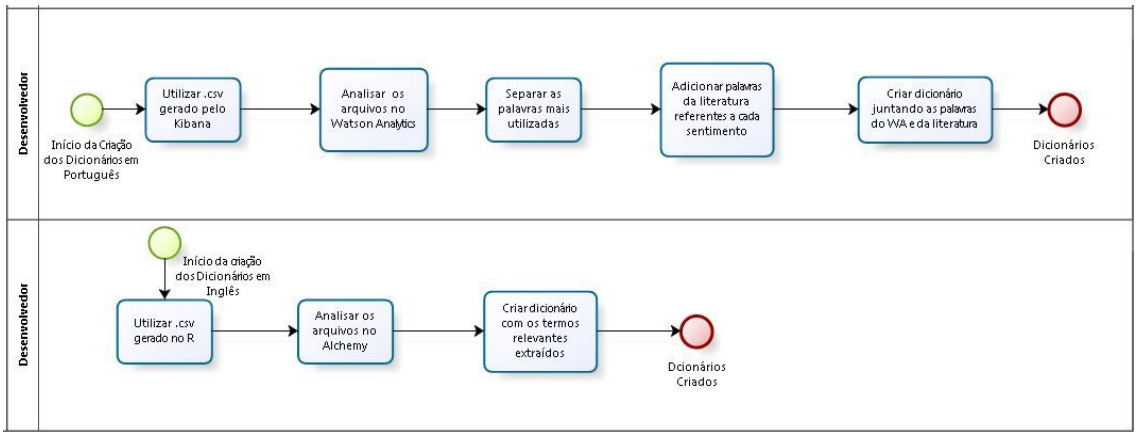

Figura 23- Processo de criação dos dicionários com o Alchemy

5.3 Coleta de tweets

A coleta dos tweets em tempo real foi diretamente afetada com o começo das olimpíadas. No período anterior aos jogos, levava em torno de 30 minutos para capturar em torno de 500 tweets.

O primeiro pico de tweets ocorreu no primeiro primeiro dia que ocorreu um jogo de futebol. Assim que o pacote ELK foi iniciado e a interface do Kibana carregada no navegador, instantaneamente, carregaram em torno de 1000 tweets com a hashtag #Rio 2016. Esse padrão se manteve até o final dos jogos. Após os jogos, o tempo de captura voltou a aumentar, mas ainda se manteve menor que o tempo de captura anterior aos jogos.

Esse padrão pode indicar que, apesar de as pessoas saberem o que estava acontecendo em relação às olimpíadas, elas se manifestaram muito mais durante o período dos jogos do que antes ou depois.

5.4 Uso do aplicativo Alchemy

A extração de palavras-chave utilizando o Alchemy gerou os dicionários positivo e negativo utilizados para classificar os tweets em inglês nos três períodos analisados. O Alchemy classifica as palavras-chave com sentimentos positivo e negativo e essa classificação foi usada para definir quais palavras estariam em cada dicionário.

O demo online do Alchemy analisava em torno de 500 tweets por vez. Uma lista com as palavras-chave separadas por sentimento foi criada e algumas palavras da lista foram selecionadas para montar os dicionários, pois a lista continha palavras repetidas como olympics e Olympics com o mesmo sentimento, mas para o R a letra maiúscula ou minúscula não fazem diferença.

Algumas palavras foram classificadas como positivas e como negativas pelo Alchemy. Elas foram incluídas nos dois dicionários. O fato de as palavras terem sido classificadas nos dois sentimentos indica que existiam opiniões diferentes sobre o mesmo assunto. Palavras ligadas às medalhas, como gold e bronze, são exemplos de palavras que estavam presentes em tweets positivos e negativos, ou seja, enquanto para alguns usuários a medalha era algo valioso para outros nem a de ouro era algo bom ou os usuários estavam reclamando que outro país havia ganhado o ouro.

A Tabela 1 mostra as palavras que foram usadas em cada um dos dicionários em inglês em cada período analisado.

Tabela 1- Dicionários em inglês

Antes	Palavras
Positivo	flag_bearer, roadtorio, museum, opening_ceremony, athletes, luck, medalist, rings, refugees, beach_volleyball, performance, summer, medal, village, volunteers, moments, vinicius
Negativo	explosive, village, risk, water, soccer, ceremony, zika, shame, olympics, terror, islamic, terrorism, fever
Durante	
Positivo	village, athletes, performance, ceremony, torch, warriors, nbcolympics, gymnastics, gold, bronze, maracana, carioca arena
Negativo	competition, usa, basketball, nbcolympics, haters, carmelo, hokey, tennins, rain, lotche, bronze
Depois	
Positivo	USA, gold, Usain, Brasil, Congrats, Gabriel, medal, game, ceremony, Biles, moment, volunteers, Paralympics, fair, lochte, heroics
Negativo	gold, lee, penalty, comments, medal, olympics, cops, robbery, scandal, champion, football

No período anterior às Olimpíadas, palavras como explosivos, zika, terror e terrorismo apareceram como palavra-chave, indicando que as pessoas de fora do país estavam preocupadas com terrorismo e possíveis surtos de doença como a zika.

Durante as olimpíadas, tanto o dicionário positivo quanto o negativo continham apenas palavras relacionadas aos jogos mostrando que as preocupações que existiam antes das olimpíadas começarem já não eram tão comentadas.

O dicionário após olimpíadas mostrou palavras como “cops”, “robbery” e “scandal” aparecendo como negativas e “Lochte” aparecendo como positiva pode indicar que algumas pessoas acreditavam na história contada pelo nadador enquanto outras pessoas não aprovaram o episódio que ocorreu no posto na Barra.

A palavra “volunteer” foi uma das que apareceu apenas com sentimento positivo, mostrando que possivelmente as pessoas gostaram do trabalho dos voluntários durante os jogos. De fato, era fácil encontrar voluntário bem humorados, que sabiam dar informações e falando outros idiomas.

Nomes de atletas como Simone Biles e Gabriel Jesus também apareceram com sentimento positivo após as Olimpíadas, isso pode indicar que a performance dos dois atletas agradou o público dos jogos.

As palavras “penalty” e “football” apareceram como negativas após os jogos, um indicativo de que, quem não era brasileiro, pode não ter gostado do resultado final do futebol, em que o Brasil ganhou o ouro inédito.

Uma das maiores polêmicas antes do início dos jogos foi o fato de a vila dos atletas não estar pronta quando os primeiros atletas chegaram ao Rio. A palavra “village” aparece com sentimento negativo e positivo para o período anterior às Olimpíadas, apenas positivo durante as Olimpíadas e já nem apareceu no período após os jogos, mostrando que a preocupação inicial passou por completo aos longo dos períodos analisados.

5.5 Uso do aplicativo Watson Analytics

O Watson Analytics foi utilizado na criação dos dicionários em português. Os arquivos .csv gerados pelo Kibana a cada dia foram organizados em 3: um para antes das olimpíadas, um para durante e um para depois. Cada um desses 3 arquivos foi carregado para o Watson e para que a identificação dos termos mais frequentes fosse feita.

A Tabela 2 mostra as palavras que foram usadas nos dicionários em português. Algumas das palavras apareciam repetidas nos dicionários em dias diferentes e em períodos diferentes.

Tabela 2- Dicionários em português

Antes	Palavras
Positivo	cobertura, participantes, boa, feliz, sorte, ouro, bronze, medalha, apoio, gol, sonho, copacabana, cidade, felicidade, rio, atletas, orgulho, tocha
Negativo	ruim, triste, tiros, guanabara, cidade, problemas, crise, hospitais, zika, trânsito, vila, seguro, defesa, vergonha, mal, bandidos, terrorismo, rio
Durante	
Positivo	abertura, gisele, 14bis, feliz, show, demais, samba, maravilhosa, orgulho, sensacional, bem, espetacular, amei, bonito, amo, legal, melhor, festa, bronze, bom, emocionante, ouro, sonho, empoderamento, conseguimos, amor
Negativo	chora, revolta, negar, porra, mal, caralho, reclama, foratemer, nem, gambiarra, merda, golpista, socorro, chocada, medo
Depois	
Positivo	atletas, despedida, orgulhosos, sensacional, superar, curtindo, saudade, show, especial, feliz, bom, boa, legal, emocinante
Negativo	lochte, presidente, dilma, discursos, eleicoes2016, favela, ruim, triste, mal, estado

Para os tweets em português, palavras como tiros, problemas, Guanabara, hospitais, trânsito e vila indicam que o povo brasileiro estava preocupado se a cidade do Rio conseguiria comportar o evento e se tudo estaria pronto a tempo e não apenas com terrorismo e os possíveis surtos de zika.

A palavra “Lochte” apareceu como negativa nos dicionários em português mostrando que o episódio no posto de gasolina[4] uma repercussão ruim para brasileiros e que, pelo menos os usuários do Twitter, não acreditavam na história dos nadadores.

Palavras como “foratemer”, “Dilma”, “golpista” são exemplos que surgiram durante e após as olimpíadas, ou seja, mesmo com foco da imprensa totalmente focado nos jogos, a situação política no país não estava completamente esquecida pelo povo.

As palavras do dicionário de durante em português não foram tão voltadas aos esportes dos jogos quanto em inglês, mas algumas palavras que apareceram eram claramente específicas da abertura como “14bis”, “Gisele”, “abertura”. A Cerimônia de abertura foi, com certeza, muito além do que todos imaginavam[5] conseguindo abordar a cultura, a música e os feitos históricos do Brasil.

[4]http://g1.globo.com/rio-de-janeiro/olimpiadas/rio2016/noticia/2016/08/video-do-posto-de-gasolina-mostra-confusao-com-nadadores-americanos.html

[5] http://veja.abril.com.br/mundo/festa-da-abertura-da-rio-2016-empolga-imprensa-internacional/

Uma diferença que pôde ser observada entre os dicionários em inglês e português é a quantidade de palavrões que aparecem. Vários exemplos foram encontrados no português e nenhum no inglês.

O dicionário após os jogos mostrou que, apesar de todos os problemas e desconfianças iniciais, de maneira geral, o brasileiro aprovou os jogos olímpicos e se orgulhou do que foi evento.

Durante as olimpíadas, em todos os períodos, foi possível observar que enquanto os estrangeiros mantinham o foco nos jogos, os elogios e reclamações dos brasileiros eram mais abrangentes, pois envolviam também os problemas comum do dia-a-dia do país.

5.6 Métodos do R

Dois métodos foram criados no R para o projeto: (i) para verificar quantas palavras dos dicionários estavam presente em cada tweet e (ii) para contabilizar quantos tweets de cada categoria de sentimento (positivo, negativo, neutro, positivo e negativo) estavam presentes em cada período de tempo analisado.

O método que verifica a quantidade de palavras dos dicionários nos tweets recebe três objetos como parâmetro: uma DTM completa contendo todos os tweets como linhas e todas as palavras como colunas, uma DTM de comparação com um dicionário contendo todos os tweets como linhas e as palavras do dicionário como coluna e um dataframe com os tweets. A Figura 24 mostra o código do método wishTweet.

```
wishTweet <- function(df_rio_all, df_analysis, df_tweets){
        wish_Vector <- vector()
        percente_vector <- vector()
        percente <- 0
        word.count <- 0
        tweetWords <- 0
        for(i in 1:nrow(df_rio_all)) {
                tweetWords <- 0
                for(j in 1:ncol(df_rio_all)){
                        words <- df_rio_all[i,j]
                        if(words > 0){
                                tweetWords<-tweetWords+1
                        }
                }
                word.count <- 0
                for(l in 1:ncol(df_analysis)){
                        is.wish <- df_analysis[i,l]
                        if(is.wish > 0){
                                word.count<-word.count+1
                        }
                }
                if(word.count > 0){
                        percente <- (word.count/tweetWords)*100
                        percente <- format(round(percente, 2), nsmall = 2)
                        percente_vector <- c(percente_vector, percente)
                        wish_Vector <- c(wish_Vector, df_tweets[i,1])
                }else{
                        percente <- 0
                        percente <- format(round(percente, 2), nsmall = 2)
                        percente_vector <- c(percente_vector, percente)
                        wish_Vector <- c(wish_Vector, df_tweets[i,1])
                }
        }
        wishTweet_list <- data.frame(wish_Vector,percente_vector)
        colnames(wishTweet_list) <- c("Tweet","dic_Name")
        return(wishTweet_list)
}
```

Figura 24- Código do método wishTweet

O primeiro passo é contar quantas palavras um tweet tem utilizando a DTM completa. Em seguida, conta-se quantas palavras do dicionário estão no tweet, se tiver 1 ou mais a porcentagem do sentimento naquele tweet é calculada. O último passo é criar um novo dataframe que contenha os tweets e a porcentagem naquele tweet do sentimento analisado.

O método que verifica quantos tweets existem em cada categoria de sentimento recebe o dataframe que contém os tweets e a porcentagem dos sentimentos como parâmetro. Para cada tweet do dataframe, verifica-se se a porcentagem de cada sentimento e a separação deles nas 4 categorias. A Figura 25 mostra o código countTweets.

```
countTweets <- function(df_final){
        positive <- 0;
        negative <- 0;
        both <- 0;
        none <- 0;
        count <- 0;
        feeling <- 0;

        for(i in 1:nrow(df_final)) {
                for(j in 2:ncol(df_final)){
                        feeling <- as.numeric(as.character(df_final[i,j]))
                        if(feeling > 0){#if true, tweet is positive
                                feeling <- as.numeric(as.character(df_final[i,j+1]))
                                if(feeling > 0){#if true, tweet is positive and negative
                                        both<-both+1
                                }else{#tweet is only positive
                                        positive<-positive+1
                                }
                                break
                        }else{
                                feeling <- as.numeric(as.character(df_final[i,j+1]))
                                if(feeling > 0){#if true, tweet is only negative
                                        negative<-negative+1
                                }else{#tweet is not positive nor negative
                                        none<-none+1
                                }
                                break
                        }
                }
        }
        cat("positive ",positive)
        cat("negative ",negative)
        cat("both ",both)
        cat("none ",none)

}
```

Figura 25- Código do método countTweet

5.7 Sentimentos ao longo dos períodos

A classificação dos sentimentos para os tweets em inglês mostrou que no período anterior às olimpíadas, quase 74% dos tweets apresentava sentimento positivo enquanto apenas cerca de 1% apresentavam sentimento negativo. Aproximadamente 21% dos tweets continham tanto sentimento negativo quanto positivo e 4% dos tweets eram neutros. A Figura 26 mostra o gráfico que representa a porcentagem de cada uma das categorias de sentimentos no período anterior às olimpíadas.

Tweets como “@CBCSask: 1 month to go til #Rio2016! RT if you're ready to cheer on #ourathlete @btheiseneaton”, “@CalAthletics: Cause I got a really big team! #GoBears #Rio2016”, “Can't wait to see @usainbolt fly - sheer magnificence #RioOlympics #Rio2016” são exemplos de tweets positivos antes dos Jogos Olímpicos. Eles mostram que os usuários estavam felizes em demonstrar seu apoio aos seus atletas e/ou times preferidos.

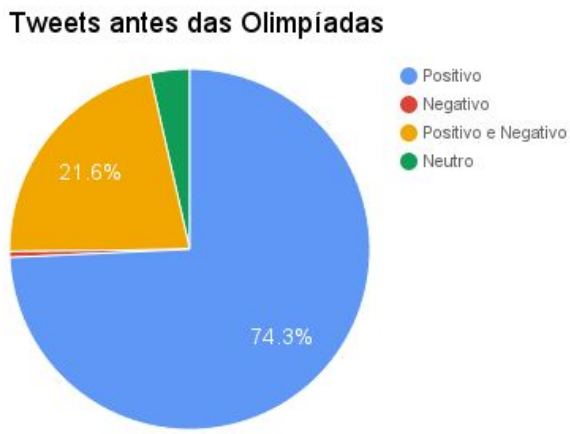

Figura 26- Gráfico de sentimento antes das Olimpíadas

Durante as olimpíadas, os sentimentos mudaram. Os tweets neutros aumentaram e chegaram a quase 61% e a quantidade de tweets negativos superou a de positivos. Esse aumento nos tweets negativos pode indicar que os estrangeiros não estavam felizes com o desempenho dos seus países nos jogos. A Figura 27 mostra o gráfico que representa a porcentagem de cada uma das categorias de sentimentos no período durante as olimpíadas.

Tweets como “@WNBA: #USA advances to the #Rio2016 Final (vs. #ESP on Saturday), defeats #FRA 86-67!”, “@TheHockeyIndia: Hockey India congratulates the German Men's #Hockey Team for bagging the Bronze Medal at #Rio2016 on 18 Aug.” mostram a torcida para os times e sem perder o espírito esportivo, como esse tweet do time indiano de Hockey parabenizando o time alemão mesmo tendo perdido para eles na etapa inicial dos jogos.

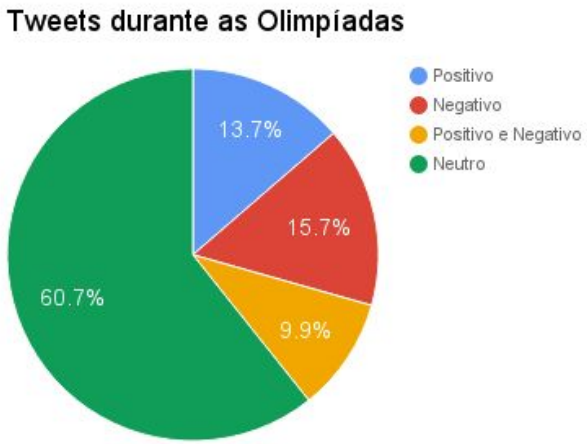

Figura 27- Gráfico de sentimento durante as Olimpíadas

Após as olimpíadas, os tweets neutros permaneceram como maioria, mas o número de tweets que mistura os sentimentos positivo e negativo superou o número de tweets apenas positivos e apenas negativos. Mais uma vez o sentimento negativo esteve presente em mais tweets do que o sentimento positivo. A Figura 28 mostra o gráfico que representa a porcentagem de cada uma das categorias de sentimentos no período após olimpíadas.

Alguns dos tweets negativos eram referentes a performances inesperadas nas olimpíadas, mas para o lado ruim, como mostra esse tweet da NBCOlympics “2012 champion Jordan Burroughs fights back tears talking about his disappointing #Rio2016.”

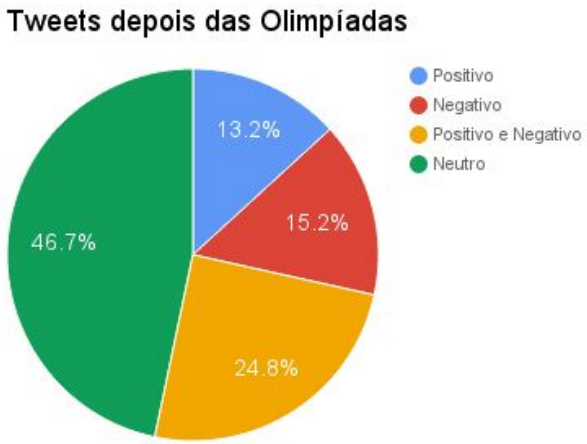

Figura 28- Gráfico de sentimento após as Olimpíadas

A Figura 29 mostra o gráfico para cada um dos sentimentos ao longo dos períodos. Neste gráfico é possível observar que o sentimento positivo era o mais presente antes das olimpíadas mas foi perdendo presença e terminou como o sentimento menos identificado nos tweets.

A categoria positivo e negativo era a segunda mais observada antes dos jogos, durante os jogos foi a menos detectada e após os jogos foi, novamente, a segunda categoria mais identificada.

Os tweets classificados como negativos foram os menos identificados antes das olimpíadas, aumentaram durante as olimpíada e diminuíram novamente após os jogos.

O tweets neutros não apareceram tanto antes das olimpíadas, mas os mais frequentes durante e após os jogos.

A maior parte dos tweets após os jogos falavam das medalhas dos últimos dias de competição. Tweets como “@Spark_Sports_: #Olympics: #Serbia demolishes #Australia 87-61, will verse #USA in #Gold matchup Sunday. #Rio2016”, “@BBCSport: You've done yourself proud #TeamGB at #Rio2016, there could me more golds yet!” são exemplos de tweets que tem a palavra gold e portanto ficam na categoria positivo e negativo ao mesmo tempo, pois a palavra gold estava atrelada aos dois sentimentos. Esse pode ser um dos motivos de os tweets com os sentimentos individuais diminuirem em quantidade ao final da análise.

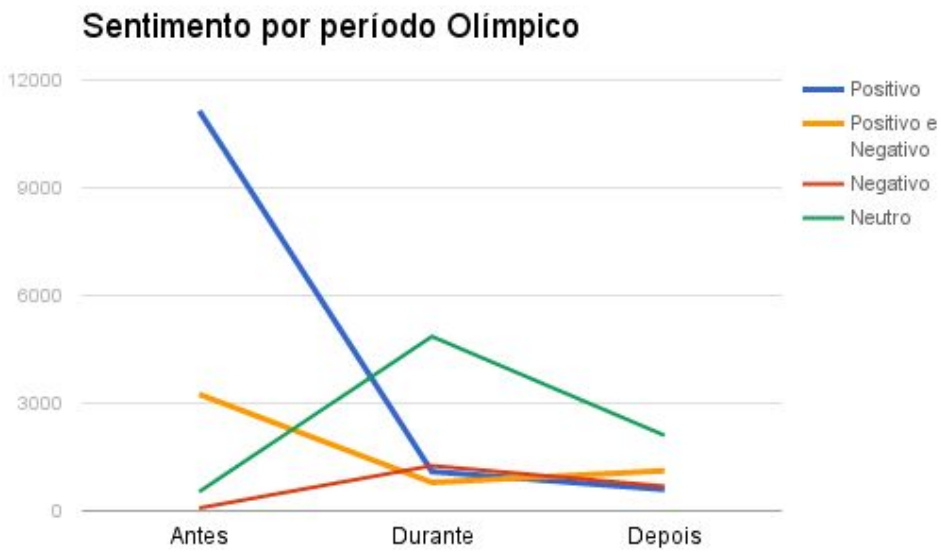

Figura 29- Gráfico dos sentimentos ao longo dos períodos

Para os tweets em português, a classificação foi um pouco mais complicada do que em inglês. O fato de a palavra “#Rio2016” ter sido usada pelos brasileiros para os mais diversos assuntos fez com que a classificação dos tweets não fosse tão eficiente quanto a dos tweets em inglês.

A quantidade de palavras nos dicionários foi aumentada, quase o dobro da quantidade nos dicionários em inglês, em uma tentativa de melhorar a classificação. Mesmo com a mudança nos dicionários, a maioria dos tweets continuava fazendo parte da categoria neutro.

Uma classificação manual poderia ser feita para definir o sentimento de cada um dos tweets. A Figura 30 mostra os gráficos obtidos para os sentimentos antes durante e depois das olimpíadas. Durante os três períodos é possível observar que a quantidade de tweets neutros, ou seja, que não contém nenhuma palavra de nenhum dos dicionários é maior que 80%.

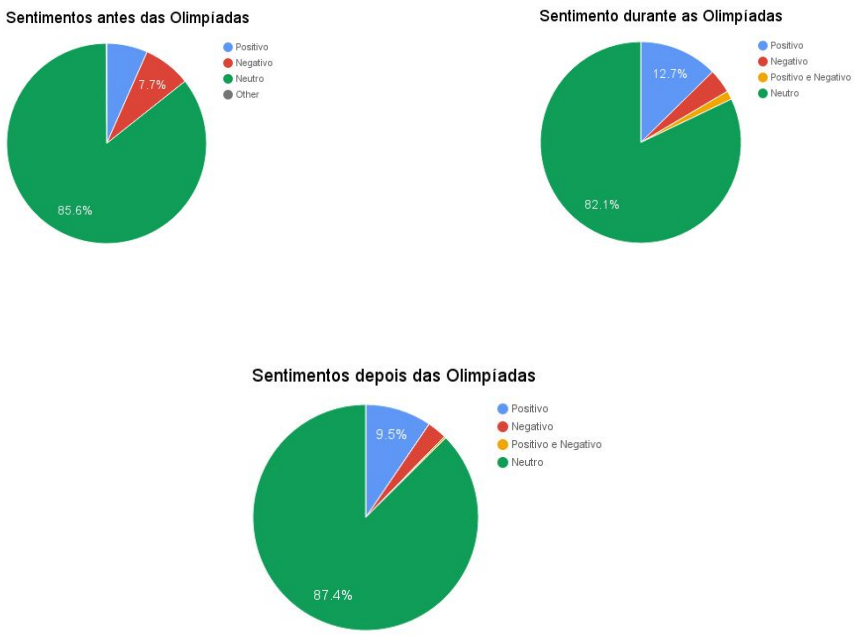

Figura 30- Gráficos para os sentimentos antes, durante e depois das Olimpíadas

Tweets falando de trânsito “@IndignadoRJ: Faltam 30 dias para as Olimpíadas. O trânsito está muito pior do que antes e as ruas esburacadas como nunca. #Rio2016”, política “#Lulanacadeia vs. #Moro o povo é Moro. #Molusconopuçá. Td espécime d Criminoso julga-se acima d LEI. #Basta #Rio2016”, do prefeito “Que cidade precisa de inimigos quando se tem um prefeito abobalhado que fala asneira toda hora como @eduardopaes_ ? #Rio2016”, violência “Rio de Janeiro #Rio2016 Helicóptero da Globo leva saraivada de tiros. #Terrorismo”, entre outros, podem ser encontrados usando a hashtag utilizada na análise.

5.8 Tweets

A Tabela 3 mostra exemplos de tweets de cada categoria de sentimento em cada um dos períodos analisados. Os exemplos de tweets negativo e positivo são bem claros no sentimento que passam, já o neutro e o que contém sentimentos positivo e negativo ao mesmo tempo são menos claros no sentimento.

Para o “antes” fica claro que o usuário que postou o tweet ficou feliz com a escolha Anna Meares como porta bandeira da Austrália. Já o usuário do tweet classificado com ambos os sentimentos fala coisas ruins a respeito da cidade, mas ainda assim escreve “Welcome to Rio”, que deveria ser algo positivo. Os tweets

neutros, pelo menos a maioria, eram tweets de notícias, não tinham efetivamente a opinião de um usuário. O exemplo negativo é bem claro também com relação ao sentimento ao afirmar que o possível surto de zika faria com que alguns atletas não viessem aos jogos.

O “durante” positivo teve como exemplo um tweet comemorando a medalha de ouro pro Canadá. Já o tweet com os dois sentimentos fala de uma derrota da França e uma vitória dos Estados Unidos. O neutro apenas se refere a um link e o negativo fala do caso dos nadadores que arrumaram confusão no posto de gasolina.

O “após” positivo parabeniza o time de basquete americano que ganhou o ouro nos jogos. O tweet com os dois sentimentos fala bem da cobertura feita no Facebook enquanto reclama da cobertura feita na tv a cabo americana. O tweet neutro fala de uma promoção e como participar dela. O tweet negativo lamenta por bons atletas que poderiam ter ganhado medalhas, mas não ganharam.

Tabela 3- Exemplos de tweets em inglês

Antes	Tweets
positivo	Congrats @AnnaMeares on being selected as flag_bearer for @Rio2016
positivo e negativo	Body parts on the beach? Check. Zika virus? Check. Disgruntled police? Check. Super resistant bacteria? Check. Welcome to #Rio2016 #Olympics
neutro	RT @MonishNand: Fiji 7s team going thru their training run this morning. The team to @Rio2016_en will be announced next week. https://t.co/…
negativo	RT @boba_oudou: Many #golfers will be #excused from #Olympics @Rio2016_en due to risks of #Zika virus infection. #ジカ熱 #オリンピック #nhk11 https:…
Durante	
positivo	RT @CBC: Erica Wiebe is an Olympic Champion in #wrestling! https://t.co/CjtnLfVlyr Another #gold for #CAN at #Rio2016 https://t.co/dCVTPuPl…
positivo e negativo	#USA women's #basketball defeat France, will battle Spain for #gold. #Rio2016 https://t.co/4vMcDmqc98 https://t.co/GbUCyQ0erl
neutro	RT @philosopop: When an American tries to out-malander a Brazilian malander #Rio2016 https://t.co/KUQ8BGXBYb
negativo	RT @RollingStone: See footage of Ryan Lochte and #Rio2016 U.S. swimmers' alleged robbery https://t.co/O87YR6q0vX https://t.co/oymeV3n8tb
Depois	
positivo	RT @SummerSanders_: Congrats on your #GoldenMoment @usabasketball men! #Rio2016 Enjoy!! #TeamUSA https://t.co/s8WGKED76J
positivo e negativo	Facebook's Olympics Performance Grabs Gold Over Comcast #media #mediawatch #Facebook #Olympics2016 #Rio2016 https://t.co/PTlWqRWAaI
neutro	Apple Mac Macbook Laptop Notebook Snow white OSX PC iTunes win : rt & follow #Rio2016
negative	9 People That Could Have Won Medals For Nigeria At The #Rio2016 Olympics. Cc @I_pissVodka https://t.co/YPb370PJ2C (via @feyinemag),

A Tabela 4 mostra exemplos de tweets em português para cada sentimento em cada um dos períodos analisados. Os exemplos em português, assim como em inglês, são bem claros para os sentimentos positivo e negativo, mas não tanto para o neutro e o positivo e negativo juntos. Para cada um dos sentimentos e períodos analisados foi colocado um exemplo de tweet relacionado aos jogos e um que não era relacionado. vct

A grande diferença entre os tweets nos dois idiomas é o foco no uso da palavra "#Rio2016". Nos tweets em inglês a grande parte tratava realmente de assuntos voltados para os jogos olímpicos. Em português, existem tweets tratando dos jogos, mas também de política, dos problemas do dia-a-dia dos brasileiros (trânsito, por exemplo), entre outros. Enquanto os tweets em inglês mostravam preocupação com surtos de Zika ("@Michael_Lipin: Can #Rio2016 organizers overcome #Zika #doping challenges with 1month 2 go?"), brasileiros se preocupavam com a possibilidade da estrutura da cidade não aguentar um evento como os Jogos Olímpicos ("Faltam 30 dias para as Olimpíadas. E os hospitais continuam caindo aos pedaços como sempre. #Rio2016"). As informações que os estrangeiros tinham sobre os possíveis problemas da cidade não eram completas. Eles tinham noção de uma parte crítica: possíveis surtos de doença, mas tinham muitos outros problemas que não eram mostrados na mídia nacional ou internacional, mas que o carioca presenciava todo dia na rotina dele.

O "antes" positivo mostrou um exemplo de tweet que estava feliz com o sonho de ter uma Olimpíada aqui e um tweet falando que Copacabana é um bairro bonito. Para o positivo e negativo, um tweet falava do museu Cidade Olímpica e o outro falando da segurança na cidade do Rio sem acreditar que ela estaria garantida como afirmou o ministro. Um dos neutros falava de política, apoiando a prisão do Lula e um apenas informava que a força nacional começaria a agir no Rio. Nos negativos, um reclama do trânsito no Rio, enquanto o outro não acreditava que os jogos olímpicos seriam bem sucedidos no Rio.

O "durante" positivo mostra um tweet que descreve a sensação única de ganhar uma medalha e um tweet que, mais uma vez, fala de política. Os exemplos positivo e negativo falam tanto de política quanto de Olimpíada no mesmo tweet. O primeiro neutro, mais uma vez, fazia menção ao cenário político do Brasil e à Olimpíada ao mesmo tempo, o segundo falava de torcidas de futebol. Já o negativo, um é uma reclamação do nível do futebol masculino e o outro falava apenas de política, contra o Presidente Interino Michel Temer.

O "após" positivo fala de amizade nos jogos e elogia o parque olímpico. Um dos tweets classificados como positivo e negativo afirma que sentirá falta dos jogos enquanto o outro dá crédito ao governo Dilma e Lula pelo apoio a alguns atletas olímpicos. O primeiro neutro fala sobre uma disputa de quem vai sentar no banco da frente de um carro e o segundo fala sobre resultados do basquete. Um dos negativos aborda o caso dos nadadores, mais uma vez comentado pelos usuários, e o outro

desmerece os movimentos feministas e a discussão em torno da legalização do aborto.

Tabela 4- Exemplos de tweet em português

Antes	Tweets
positivo	RT @Rafaelasilvaa: Falta apenas UM mês para o início das Olimpíadas Rio 2016. UM mês para começar a realizar o sonho! #Rio2016 https://t.co… Copacabana quanta beleza nas férias de julho #rio2016 #valeRio2016 #copacabana #posto5 #streetview #riodejaneiro... https://t.co/Xj5sAomtXO
positivo e negativo	Museu Cidade Olímpica convida visitantes a sobrevoar o Rio de ... - Rio 2016 https://t.co/ClwTbrRrcm "A um mês da #Rio2016 ministro da justiça diz que segurança está garantida no Rio" kkkkkkkkkk me sinto beeem mais seguro depois dessa. SQN
neutro	#Lulanacadeia vs. #Moro o povo é #Moro #Lava-Jato td @ptbrasil preso Vaccari&cia @ECantanhede #Lulaladrao #WelcomeToHell #Rio2016 @cnni RT @ebcnarede: Força Nacional assume hoje controle das áreas de competição da @Rio2016 https://t.co/hKuAO03lmH https://t.co/kYiHhDCLWO
negativo	#rio2016 #eventocatastrófico #whatashame Quando passar vergonha é o esporte favorito! https://t.co/0qAM8Ec5y5 to passando mal e to no engarrafamento #adoro #rio2016
Durante	
positivo	RT @ribasdiego10: Disputar os Jogos Olímpicos e ganhar uma medalha é uma sensação única. Boa sorte à Seleção! #Rio2016 #bra #futebol https:… RT @PanichiRaphael: Essa é a medalha de Ouro que o Brasil quer! Lula na Cadeia! Por isso #EuGritoMoro nas Olimpíadas #Rio2016 https://t.co/…
positivo e negativo	RT @pahega: Já saiu a primeira medalha de ouro do Brasil!!!! Golpe Sincronizado #Rio2016 #ForaTemer https://t.co/ZnClznMKII RT @paulocoelho: Chatos: #ForaTemer sim, mas parem de criticar #Rio2016 . Boa vibração atrai boa energia e pode mudar o país.
neutro	RT @AngelaMaximo1: #MoroE10Medidas MoooooRooooo estreando na Abertura da #Rio2016 https://t.co/Cs1nh2Vkly RT @donas_da_bola: A torcida de Portugal! Não, pera! #Rio2016 #Futebol https://t.co/gGaev4OWaU
negativo	O futebol brasileiro começou mal na olimpíada ⚽ #Futebol #olympics #Rio2016 #Rio2016 Grita forte Brasil #ForaTemer https://t.co/SAPRL4aLJM
Depois	
positivo	RT @McDonalds_BR: A #AmizadeVence nos Jogos #Rio2016! Estamos orgulhosos de nossas crianças. Medalhistas de ouro na categoria amizade! http… RT @RedeGlobo: Boa noite, parque olímpico! Uma visão exclusiva do nosso estúdio. #SomosTodosOlímpicos #Rio2016 https://t.co/W1cX6ky8R5
positivo e negativo	RT @alamuss: Triste pelo fim da Olimpíada mas muito feliz por ter feito parte dela. As memórias são eternas. #Rio2016 RT @esuplicy: A maioria dos medalhistas recebeu incentivos criados nos governos Lula e Dilma: Segundo Tempo, Atletas de Rendimento e Bolsa…
neutro	Eu e minha irmã pra ver quem vai no banco da frente do carro #Rio2016 https://t.co/balwmLKhte #Rio2016 || #Basquetebol A Sérvia derrota a Austrália por 87x61 e enfrenta #USA na disputa do #ouro
negativo	Comitê Olímpico dos EUA deve estar puto, pois, todo as medalhas que ganharam foram ofuscadas pelo caso Ryan Lochte. #Rio2016 RT @HumanasNao: Feminista chora Abortista passa mal #Rio2016 https://t.co/pk9uukGUrA

6 Conclusão

Este trabalho apresentou uma análise de sentimentos relacionados aos Jogos Olímpicos realizados em 2016 no Rio de Janeiro. Foram analisados tweets em dois idiomas - inglês e português - e utilizando métodos diferentes para cada idioma.

Duas formas de captura de tweets foram abordadas: a captura em tempo real e sem ser em tempo real. A captura em tempo real normalmente trazia tweets relacionados ao mesmo assunto, algum jogo ou acontecimento recente, enquanto a captura em tempo não real trazia uma mistura de assuntos nos tweets ao longo do dia. Para a captura foram usados o software R e o pacote ELK (Elasticsearch, Logstash e Kibana). As ferramentas escolhidas não são as únicas opções disponíveis para realizar esse tipo de trabalho, elas foram escolhidas por estarem muito bem documentadas e por interesse no aprendizado delas especificamente. Um exemplo de análise dos jogos, especificamente da abertura, foi feita pela empresa Gorkana e utilizando ferramentas próprias [36].

A criação dos dicionários para os sentimentos foi criada de maneira específica para cada idioma. Para os dicionário em inglês, o Alchemy foi usado para fazer a extração das keywords para cada um dos sentimentos e para português utilizou-se os termos significativos extraídos do ELK e analisados no Watson Analytics.

O R foi utilizado para o pré-processamento dos tweets e para analisar o sentimento contido neles. Cada tweet era comparado com cada um dos dicionários

(positivo e negativo) e a porcentagem de sentimento (positivo, negativo, neutro ou positivo e negativo) era calculada.

Para os tweets em inglês foi possível observar uma mudança nos sentimentos ao longo dos períodos analisados. Antes das olimpíadas o sentimento positivo era o mais forte e o negativo era o mais fraco. Durante as olimpíadas, o neutro foi o mais forte e o que continha sentimento positivo e negativo o mais fraco. Após as olimpíadas, o neutro continuava como mais forte enquanto o positivo era o mais fraco.

A classificação dos tweets em português não foi tão eficiente quanto em inglês. Um dos motivos que pode ter diminuído a eficiência da classificação é o fato de que os brasileiros tuitavam sobre os mais diversos assuntos usando “#Rio2016” e não apenas para assuntos envolvendo as olimpíadas.

A criação de modelos de processo para cada uma das etapas do projeto pode ser útil para os próximos passos da pesquisa. Com um esforço não muito grande um workflow científico pode ser implementado para apoiar a realização destes experimentos.

Para que essa proposta seja utilizada por uma empresa ou um pesquisador da área seria necessário, além de seguir os diagramas propostos, uma etapa de revisão dos dicionários por um especialista no assunto analisado.

Como trabalhos futuros do projeto, pode ser proposta uma mudança na metodologia da análise para tweets em português, possivelmente usar a versão paga do Alchemy para ter acesso a extração de palavras-chave em português e reanalisar os tweets. É possível que a extração de keywords que já tenham sentimento definido melhore a separação dos tweets e torne a análise mais eficiente.

Uma classificação manual prévia para a criação dos dicionários poderia ajudar considerando os diferentes assuntos abordados nos tweets em português. Ou uma separação, mantendo apenas os tweets relacionados efetivamente às olimpíadas poderiam tornar a classificação mais eficiente.

Referências Bibliográficas

1. Baingana, B., Traganitis, P., Giannakis, G., Mateos, G., (2016) “Big Data Analytics for Social Network”, Graph-Based Social Media Analysis, CRC Press
2. Chen, H., Chiang, R. H. L., Storey, V. C., (2012) “Business Intelligence and Analytics from Big Data to Big Impact”.
3. Wlodarczak P., Soar J., Ally M., Big Data of Social Media, Recent Advances in Computer Science http://www.inase.org/library/2015/zakynthos/bypaper/COMPUTERS/COMPUTERS-02.pdf
4. Bollen J. , Mao H. , and Zeng X.-J. , Twitter mood predicts the stock market, Journal of Computational Science, vol. 2, pp. 8, 2010.
5. Asur S. , and Huberman B. A. , Predicting the Future with Social Media, presented at the IEEE Int. Conf. Web Intelligence, 2010, pp. 492-499.
6. Achrekar H. , Gandhe A. , Lazarus R. , Ssu-Hsin Y. , and Benyuan L. , Predicting Flu Trends using Twitter data, presented at the IEEE Computer Communications Workshops (INFOCOM WKSHPS), 2011, pp. 702-707.
7. Liu, B., (2012), Sentiment Analysis and Opinion Mining, http://www.morganclaypool.com/doi/abs/10.2200/s00416ed1v01y201204hlt016
8. Dumas, M., La Rosa, M., Mendling, J., Reijers, H. A., (2013) Fundamentals of Business Process Management, Springer.
9. Documentação do Scrum http://www.desenvolvimentoagil.com.br/scrum/
10. Han, J., Kamber, M., Pei, J. (2012), Data Mining: Concepts and Techniques, Elsevier, 3rd edition.
11. GSI - Grupo de sistemas inteligentes (1998), “Algoritmos de mineração”, http://www.din.uem.br/ia/mineracao/tecnologia/index.html
12. Wu, X, et al (2008), “Top 10 algorithms in data mining”, Knowledge and Information Systems 14:1–37.
13. Morais, E. A. M., Ambrósio, A. P. L, (2007), “Mineração de texto”, http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf
14. Aranha, C., Passos, E., (2006), A Tecnologia de Mineração de Textos, RESI-Revista Eletrônica de Sistemas de Informação, Nº2, http://www.periodicosibepes.org.br/ojs/index.php/reinfo/article/viewFile/171/66
15. Aggarwal, C. C., Zhai, C. X. (2012), Data text mining, Springer

16. Wives, L. K. (2002), “Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva”, Exame de qualificação EQ-069 PPGC-UFRGS
17. Agarwal, A. et al, (2011) “Sentiment Analysis of Twitter Data”, http://dl.acm.org/citation.cfm?id=2021114
18. Wasserman, S., Faust, K. (1994), Social Network Analysis: Methods and Applications, Cambridge University Press.
19. Aggarwal, C. C., (2011), “An Introduction to Social Network Data Analytics” Social Network Data Analysis, Springer
20. Pak, A., Paroubek, P. (2010) “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf
21. Russell, M. A. (2014), Mining the social web, O’Reilly Media, 2nd edition.
22. Liu, B. (2010) “Sentiment analysis and subjectivity”, Handbook of natural language processing, 2nd edition.
23. Pang, B., Lee, L. (2008), “Opinion mining and sentiment analysis”, http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf
24. Donkor, B. (2013), “On social sentiment and sentiment analysis”, http://brnrd.me/social-sentiment-sentiment-analysis/
25. Liu, B. (2012) Sentiment analysis and opinion mining, Morgan and Claypool Publishers.
26. Cavalin P. R. et al, (2014) “Real-time Sentiment Analysis in Social Media Streams: The 2013 Confederation Cup Case”, IBM Research, http://www.lbd.dcc.ufmg.br/colecoes/eniac/2014/0094.pdf
27. Wang H. et al (2012), “A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle”, http://dl.acm.org/citation.cfm?id=2390490
28. Thelwall M., Wilkinson D., Uppal S. (2009), “Data mining emotion in social network communication: Gender differences in MySpace”, http://onlinelibrary.wiley.com/doi/10.1002/asi.21180/full
29. Documentação do Twitter, https://dev.twitter.com/rest/reference/get/search/tweets
30. Documentação do R, https://www.r-project.org/
31. Zhao, Y. (2013), “R and data mining: Examples and case studies”, https://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf

32. Gentry, J., (2016), Package TwitteR, https://cran.r-project.org/web/packages/twitteR/twitteR.pdf
33. Gentry, J., (2016), Package ROAuth, https://cran.r-project.org/web/packages/ROAuth/ROAuth.pdf
34. Documentação do Elastic Stack, https://www.elastic.co/products
35. Documentação Alchemy API, http://www.ibm.com/watson/developercloud/doc/alchemylanguage/
36. http://www.gorkana.com/2016/08/rio-olympics-2016-twitter-analysis-infographic/