



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

ESCOLA DE INFORMÁTICA APLICADA

**Processo de Mineração de Desejos aplicado em dados dos Jogos
Olímpicos Rio 2016**

Gabriel Pereira Ramos

Orientador

Fernanda Araujo Baião Amorim

Coorientador

João Carlos de Almeida Rodrigues Gonçalves

RIO DE JANEIRO, RJ – BRASIL

JULHO DE 2017

Catálogo informatizada pelo(a) autor(a)

P175 Pereira Ramos, Gabriel
 Processo de Mineração de Desejos aplicado em
 dados dos Jogos Olímpicos Rio 2016 / Gabriel Pereira
 Ramos. -- Rio de Janeiro, 2017.
 63 pgs

 Orientadora: Fernanda Araujo Baião Amorim.
 Coorientadora: João Carlos Almeida Rodrigues
 Gonçalves.

 Trabalho de Conclusão de Curso (Graduação) -
 Universidade Federal do Estado do Rio de Janeiro,
 Graduação em Sistemas de Informação, 2017.

 1. Descoberta de Conhecimento em Banco de Dados.
 2. Mineração de Desejos. 3. Rio 2016. 4. Knowledge
 Discovery in Databases. 5. Data Mining. I. Araujo
 Baião Amorim, Fernanda, orient. II. Almeida
 Rodrigues Gonçalves, João Carlos, coorient. III.
 Título.

Processo de Mineração de Desejos aplicado em dados dos Jogos Olímpicos Rio 2016

Gabriel Pereira Ramos

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:

Fernanda Araujo Baião – Orientador (UNIRIO)

João Carlos de Almeida Rodrigues Gonçalves - Coorientador

Ana Cristina Bicharra Garcia

Flávia Maria Santoro

RIO DE JANEIRO, RJ – BRASIL.

JULHO DE 2017

RESUMO

O sentimento de desejo já estudado pelas ciências humanas e psicológicas ganha uma nova perspectiva de estudo nas ciências exatas. Aproveitando da oportunidade criada pelos Jogos Olímpicos de 2016, sediado no Brasil, para grande captura de dados textuais em português e inglês da rede social Twitter foi possível estudar os motivos que levam a comunidade falante do português e a comunidade falante do inglês a manifestar desejos em relação a pessoas associadas aos Jogos Olímpicos. Utilizando métodos de processamento de texto em linguagem natural, mineração de textos para encontrar os desejos, análise de sentimento para classificação de desejos e técnicas de refinamento para exposição dos desejos foi possível levantar fatores que podem motivar desejos. Foi possível observar que desejos positivos são fruto de contínuo cumprimento de expectativas geradas pelos fãs para determinado atleta e que apenas uma expectativa não cumprida pode levar a desejos negativos para um atleta por dias. Fatores de desejo não relacionado a atletas também foram observados como atenção da mídia, política e nacionalidade. Principais verbos relacionados aos desejos realizados durante os Jogos Olímpicos também foram levantados. Apesar de um desejo poder ser tão complexo quanto a mente humana foi possível desenvolver um processo que aplica técnicas computacionais avançadas de processamento de linguagem natural e de mineração de desejos para capturar indícios de desejos relacionados aos Jogos Olímpicos que pode abrir uma série de aplicações interessantes na área de Sistemas de Informação.

Palavras-chave: Descoberta de Conhecimento em Banco de Dados, Mineração de Desejos, Rio 2016

ABSTRACT

The sense of desire, already studied by the human and psychological sciences, has a new study in the exact sciences. Taking advantage of the opportunity created by the 2016 Olympic Games in Rio de Janeiro to collect large amounts of textual data in Portuguese and English from Twitter, it was possible to find reasons that lead Portuguese and English speakers to make wishes related to athletes during the Rio 2016 Olympic Games. Using natural language processing, text mining, sentiment analysis and refinement techniques it was possible to capture hints about the reasons that motivate desires. Texts with desires were classified into positive, negative or neutral. Positive desires are the result of the continuous fulfillment of expectations for an athlete and only one unfulfilled expectation can lead to negative desires for an athlete for days. We found out that media attention, politics and nationality are factors that influence wish making. Main verbs related to wishes made during the Olympic Games were also found. Although a desire can be as complex as the human mind, this study describes an approach that may open several opportunities to understand cause and effect relationship that lead to desires.

Keywords: KDD, Wish Detection, Rio 2016, Data Mining

Índice

1	Introdução	12
1.1	Motivação	12
1.2	Objetivos	14
1.3	Organização do texto	14
2	Fundamentação Teórica	15
2.1	Mineração de textos	15
2.2	Mineração de desejos em textos - o método BeliefMining	16
2.3	Naive Bayes para mineração de textos	17
2.4	Processamento da Linguagem Natural	19
2.5	Named Entity Recognizer	21
2.6	Análise de sentimentos em textos	22
2.7	Redes de coocorrência entre palavras	23
3	Processo de descoberta de desejos	25
3.1	Visão geral	25
3.2	Coleta	26
3.3	Tradução	28
3.4	Pré-processamento - técnicas PLN	28
3.5	Treinamento Naive Bayes para identificação de desejos	29
3.6	Reconhecimento de Entidades Nomeadas - NER	30
3.7	Análise de sentimentos em textos	31
4	Análise de descobertas	32

4.1 Visão de Popularidade Diária	32
4.2 Visão da Popularidade Condensada	34
4.3 Visão Desejo	35
4.4 Visão Quantitativa dos Sentimentos	36
4.5 Visão de Coocorrência	44
4.6 Visão dos Desejo por idioma	49
4.7 Visão de Resultados Interessantes	53
5 Conclusão	58
6 Referências Bibliográficas	

Índice de Tabelas

Tabela 1 - Quantidade de *tweets*, *retweets* e *replies* coletadas durante etapa de coleta para língua portuguesa.

Tabela 2 - Quantidade de *tweets*, *retweets* e *replies* coletadas durante etapa de coleta para língua inglesa.

Tabela 3 – *Rank* com quantidade de *tweets* da língua inglesa marcados como desejo para os 5 atletas mais citados por dia de competição.

Tabela 4 – *Rank* com quantidade de *tweets* da língua portuguesa marcados como desejo para os 5 atletas mais citados por dia de competição.

Tabela 5 – *Rank* com quantidade de *tweets* da língua inglesa marcados como desejo para os 5 atletas mais citados nos Jogos Olímpicos.

Tabela 6 – *Rank* com quantidade de *tweets* da língua portuguesa marcados como desejo para os 5 atletas mais citados nos Jogos Olímpicos.

Tabela 7- Tabela quantitativa da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Neymar. Baseado em *tweets* da língua portuguesa.

Tabela 8- Tabela quantitativa da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Michael Phelps. Baseado em *tweets* da língua inglesa.

Tabela 9- Tabela quantitativa da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Serena Williams. Baseado em *tweets* da língua inglesa.

Tabela 10- Tabela quantitativa da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Ryan Lochte. Baseado em *tweets* da língua inglesa.

Tabela 11- Tabela quantitativa da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Diego Hypólito. Baseado em *tweets* da língua portuguesa.

Tabela 12- Tabela quantitativa da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Marta Vieira da Silva. Baseado em *tweets* da língua portuguesa.

Tabela 13 – Tabela quantitativo da classificação positivo, negativo e neutro de *tweets* marcados como desejo para *tweets* da língua inglesa.

Tabela 14 – Tabela quantitativo da classificação positivo, negativo e neutro de *tweets* marcados como desejo para *tweets* da língua portuguesa.

Índice de Figuras

Figura 1 - Interações na KIP.

Figura 2 – Teorema de Bayes adaptado para textos.

Figura 3 – Exemplo do protocolo Penn Treebank e do *Tokenization*.

Figura 4 – Processo de descoberta de desejos.

Figura 5 – *Screenshot* da GUI do Stanford NER[5].

Figura 6 – *Screenshot* da saída do Umigon[17].

Figura 7 – Diagrama de coocorrência para verbos considerando todos os *tweets* marcados como desejo da língua inglesa.

Figura 8 – Gráfico quantitativo da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Neymar. Baseado em *tweets* da língua portuguesa.

Figura 9 – Gráfico quantitativo da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Michael Phelps. Baseado em *tweets* da língua inglesa.

Figura 10 – Gráfico quantitativo da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Serena Williams. Baseado em *tweets* da língua inglesa.

Figura 11 – Gráfico quantitativo da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Ryan Lochte. Baseado em *tweets* da língua inglesa.

Figura 12 – Gráfico quantitativo da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Diego Hypólito. Baseado em *tweets* da língua portuguesa.

Figura 13 – Gráfico quantitativo da classificação positivo, negativo e neutro de *tweets* marcados como desejo para atleta Marta Vieira da Silva. Baseado em *tweets* da língua portuguesa.

Figura 14 – Diagrama de coocorrência para atleta Ryan Lochte de 05/08/2016 até 15/08/2016.

Figura 15 – Diagrama de coocorrência para atleta Ryan Lochte de 16/08/2016 até 21/08/2016.

Figura 16 – Diagrama de coocorrência para atleta Marta Vieira da Silva de 05/08/2016 até 15/08/2016.

Figura 17 – Diagrama de coocorrência para atleta Marta Vieira da Silva de 16/08/2016 até 21/08/2016.

Figura 18 – Gráfico quantitativo da classificação positivo, negativo e neutro de *tweets* marcados como desejo para *tweets* da língua inglesa.

Figura 19 – Gráfico normalizado da classificação positivo e negativo de *tweets* marcados como desejo para *tweets* da língua inglesa.

Figura 20 – Gráfico quantitativo da classificação positivo, negativo e neutro de *tweets* marcados como desejo para *tweets* da língua portuguesa.

Figura 21 – Gráfico normalizado da classificação positivo e negativo de *tweets* marcados como desejo para *tweets* da língua portuguesa

Figura 22 – Diagrama de coocorrência para artista Anitta de 05/08/2016.

Figura 23 – Diagrama de coocorrência para candidato Donald Trump de 05/08/2016.

Figura 24 – Diagrama de coocorrência para presidente Michel Temer de 05/08/2016.

Figura 25 – Diagrama de coocorrência para narrador Galvão Bueno de 21/08/2016.

1 Introdução

1.1 Motivação

Desde o surgimento da filosofia o sentimento de desejo é analisado, discutido e classificado. Segundo Aristóteles o sentimento de desejo poderia ser classificado em três formas diferentes: *epithumia* (desejo proibido relacionado como exemplo a cobiça e apetite), *thumos* (desejo relacionado a paixão) e a *boulêsis* (desejo relacionado ao querer) [1]. Pensamento de Aristóteles que foi distinguido por Sócrates com a criação do desejo racional (movido a razão) e não racional (movido a impulsos) [2]. As discussões sobre desejo sempre estiveram em sua maioria no campo das ciências humanas, Filosofia e Psicologia, mas as ciências exatas também podem contribuir com conhecimento para essa discussão.

Com o advento das redes sociais se tornou possível capturar dados textuais de uma expressiva quantidade de pessoas simultaneamente. Dados textuais de mídias sociais costumam conter opiniões e sentimentos que são possíveis de ser identificadas e detectadas utilizando técnicas de mineração [3]. Também é possível classificar os textos quanto a sentimento positivo, neutro ou negativo utilizando algoritmos de análise léxica. Essas técnicas podem ajudar a transformar os dados em conhecimento útil para situações como por exemplo decisões corporativas, tendências políticas e decisões de marketing [4].

O sentimento de desejo, já tão abordado em trabalhos de áreas humanas, Filosofia e Psicologia, pode encontrar novas análises e classificações quando utilizado em conjunto com os dados extraídos de redes sociais. Com a utilização de técnicas de processamento de linguagem natural (NLP) [5], mineração de textos [3] e análise de sentimentos [6] é possível achar padrões e tendências sobre o sentimento de desejo não apenas de uma pessoa, mas o desejo de uma comunidade.

O sentimento de desejo, na maioria dos casos, vem atrelado a um assunto ou objeto alvo. Alguém que deseja algo, deseja algo em relação a algo material ou não material. O desejo pode ser relacionado a um objeto alvo não material (como por exemplo paz mundial ou amor), ou um desejo pode ser relacionado a um objeto alvo material (como uma pessoa, um lugar ou uma organização).

Os Jogos Olímpicos de 2016 foram sediados pelo Brasil no Rio de Janeiro. Foi a primeira edição dos Jogos Olímpicos sediados na América do Sul e grande oportunidade de geração de dados sobre diversos assuntos relacionados aos jogos. Um dos principais assuntos dos Jogos Olímpicos são os competidores. Os atletas são alvo de grande visualização e atenção em todas as fases dos Jogos.

O conjunto de atletas dos jogos olímpicos é um conjunto de pessoas que podem ser objetos alvo do sentimento de desejo. Os atletas são um conjunto que é ótima fonte de dados pois possuem características e comportamento heterogêneos. Com isso o sentimento de desejo de uma comunidade com relação a cada atleta é distinto e pode variar ao longo do tempo dependendo das características individuais, do seu comportamento e do contexto onde cada atleta está inserido. Desta forma, a exploração de técnicas de NLP, mineração de textos e análise de sentimentos neste domínio pode levantar questões e oportunidades de análise interessantes e contribuir com conhecimento útil e interessante.

1.2 Objetivos

Este trabalho tem o objetivo de explorar técnicas de processamento de linguagem natural, análise de sentimentos e mineração de texto, combinando-as em um processo de descoberta de desejos. Como objeto alvo do desejo foi escolhido um objeto material que é o conjunto de atletas que participaram dos Jogos Olímpicos de 2016. Utilizando técnicas como mineração de dados, análise de sentimentos e algoritmos de visualização será realizada uma análise individual dos principais atletas que competiram nas Olimpíadas. Essa análise tem como objetivo mapear os motivos que levaram o atleta a receber desejos positivos ou negativos. A fonte de dados para o trabalho foi a hashtag oficial dos jogos no Twitter #Rio2016 [7], considerando os idiomas Português e Inglês.

1.3 Organização do texto

O trabalho está estruturado em 5 capítulos incluindo este primeiro capítulo que tem objetivo contextualizar o tema que o trabalho está inserido.

O capítulo 2 tem o objetivo de fazer uma Revisão Bibliográfica das técnicas utilizadas neste trabalho.

O capítulo 3 tem o objetivo de apresentar o processo criado para tratamento dos dados textuais e detecção dos desejos.

O capítulo 4 tem o objetivo de criar visões, com enfoque em alguma questão de análise mais específica. Neste capítulo buscou-se levantar os motivos que levaram a um atleta ser alvo de desejos positivo ou negativo, e quais atletas receberam mais desejos.

O capítulo 5 tem o objetivo de discutir os possíveis padrões, tendências e conexões que levam a um objeto alvo, mais especificamente um atleta, a ser alvo de desejos.

2 Fundamentação Teórica

Este capítulo tem o objetivo de apresentar os principais métodos e técnicas utilizados neste trabalho.

2.1 Mineração de textos

Mineração de textos pode ser definido como um processo intensivo de conhecimento em que o usuário interage com uma coleção de documentos usando ferramentas de análise com o objetivo de extrair conhecimento pela identificação e exploração de padrões, tendências e conexões [8]. O diferencial de Mineração de textos é que a fonte de dados não é encontrada em registros estruturados de bancos de dados, mas em dados não estruturados como textos em documentos [8]. A Mineração de textos utiliza o Processamento da Linguagem Natural - PLN, na fase de pré-processamento, para aprender sobre o texto, manipulando o conteúdo do texto, extraindo entidades, detectando relacionamentos, encontrando sinônimos e realizando a desambiguação do texto através de técnicas como remoção de *stop words*, *tokenization*, *lemmatization* (*Part of Speech*) e *steming* [9].

De acordo com Feldman e Sanger [8], em nível funcional a mineração de texto possui quatro subáreas: (a) Pré-processamento, (b) Operações de núcleo da mineração, (c) Camada de apresentação, (d) Técnicas de refinamento (pós-processamento).

a) O Pré-processamento inclui todas as rotinas, processos e métodos necessários

para a preparação dos dados para as operações de núcleo da mineração. As atividades de pré-processamento convertem o formato original dos dados para um formato compreensível das operações de núcleo da mineração.

b) As Operações de núcleo da mineração compõem a parte da aplicação de algoritmos de descoberta de padrões, tendências e conexões. Distribuição, frequência e associações de palavras são fatores para descoberta dos algoritmos. Alguns algoritmos utilizam bases de treinamento para aumentar sua eficácia.

c) A Camada de apresentação é composta por ferramenta visuais como gráficos e tabelas para visualização e agrupamento de conceitos, criação de perfis e exposição de padrões, tendências e conexões.

d) Técnicas de refinamento, ou pós-processamento, são métodos que permitem filtrar informação redundante, agrupar grupos de crescimento relacionado, supressão, poda, ordenação e generalização de grupos que crie uma otimização da descoberta.

2.2 Mineração de desejos em textos - o método BeliefMining

Em um processo intensivo de conhecimento os desejos, intenções e crenças são importantes fatores para entender as decisões tomadas por agentes [10]. Desejos são fundamentais nas interações de KIP como mostra a ontologia de interações na KIP Figura 1 [10]. Um processo intensivo de conhecimento pode ser a mineração de textos com foco em extrair desejos dos textos. Desejo é um fator capaz de motivar um agente e um motivo pode ser o diferencial em uma tomada de decisão. Gerar conhecimento que possa explicar a tomada de decisão de agentes, tendo em vista a visão das organizações, é fundamental para entender as necessidades de entidades que têm relações com a organização [4]. Um método de extrair padrões e tendência de intenções e desejos de participantes em um processo intensivo de conhecimento foi proposto em [3] e utiliza mineração de desejos em mídias sociais. Este método consiste na utilização do Algoritmo Naive Bayes para usando uma base de treinamento aprender a identificar o

sentimento de desejos. A partir da identificação de textos com o sentimento de desejo marcado é aplicado a técnica de análise de sentimentos capaz de distinguir sentimentos positivos e negativos dentro do sentimento de desejo, pois sentimentos positivos ou negativos podem motivar diferentes formas de desejo.

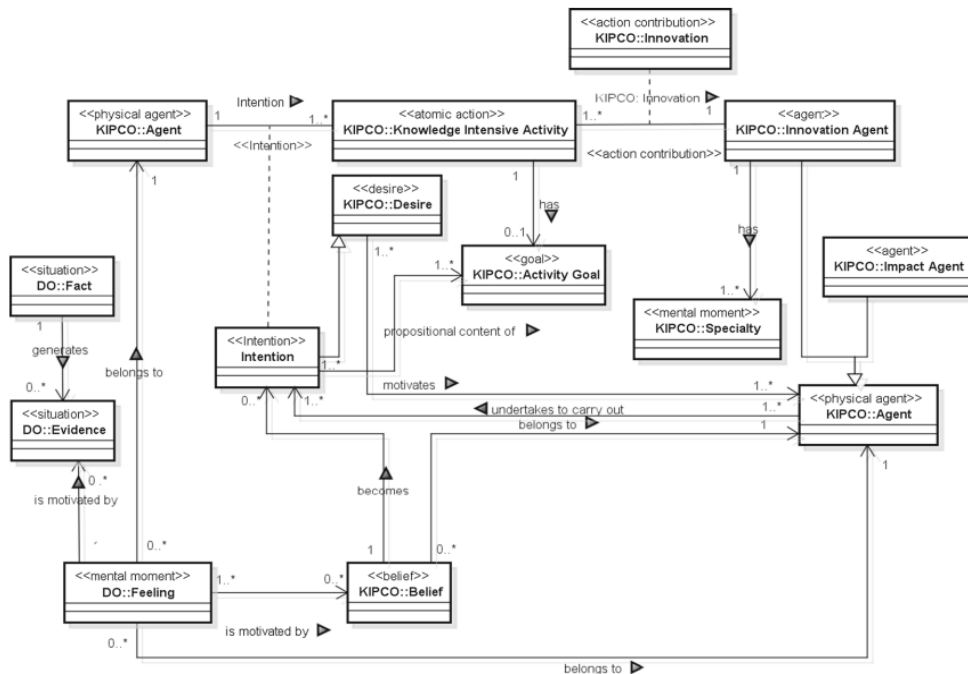


Figura 1- Interações na KIP – Fonte: FRANÇA, J., NETTO, J., CARVALHO, J., SANTORO, F., BAIÃO, F., PIMENTEL, M. et al.[10]

2.3 Naive Bayes para mineração de textos

Naive Bayes é um classificador linear conhecido por ser simples e eficiente. O modelo probabilístico do classificador é baseado no Teorema de Bayes. O adjetivo naive vem da suposição que os atributos são mutuamente independentes [11]. Para usar o Naive Bayes em textos o classificador usa o texto segmentado pela técnica de *tokenization* criando um conjunto de palavras (*bag of words*), em que cada palavra tem sua probabilidade de pertencer a uma classe. Quando aplicada a problemas de análise de

sentimentos, uma classe pode ser um sentimento como por exemplo desejo. Uma classe é calculada usando um conjunto de treinamento com textos previamente classificados [11]. A probabilidade de o texto ser de uma classe é definido pela probabilidade de cada uma de suas palavras ser da classe. O cálculo sobre o texto é feito para cada classe e a classificação é feita para a classe que obtiver maior pontuação [11]. A Figura 2 mostra o Teorema de Bayes com variáveis adaptadas para textos.

$$p(c/w) = \frac{p(w/c) * p(c)}{p(w)}$$

Figura 2 – Teorema de Bayes adaptado para textos. Fonte: RISH, Irina [11].

C é a variável para classe, W a variável para um conjunto de palavras, e P significa probabilidade. $p(W/C)$ representa a pontuação de uma palavra, dada uma classe C. Na fase de aprendizado o valor de $p(W_i/C)$, onde i é um índice para uma palavra específica, é calculado a partir do conjunto de treinamento. Quanto mais vezes a palavra W_i aparecer no treinamento para a classe C maior será sua pontuação de W_i na classe C. Quanto maior a pontuação de uma palavra para uma classe C, maior será a probabilidade de na fase de classificação de textos, os textos que contêm essa palavra serem classificados na classe C. A classificação é feita calculando a pontuação do texto, que é a soma das pontuações de cada palavra nesse texto, para cada classe [11]. A pontuação do texto em relação a classe então é comparada com a pontuação do mesmo texto para todas as outras classes. O texto é então classificado na classe que obter maior pontuação. O classificador pode utilizar técnicas de validação cruzada para gerar métricas de avaliação de qualidade da classificação na fase de treinamento [11]. Uma métrica é a precisão positiva que é a porcentagem de casos positivos classificados corretamente em relação ao total de casos classificados como positivos. Precisão negativa é a porcentagem de casos negativos classificados corretamente em relação ao total de casos classificados como negativos. A acurácia é a porcentagem de casos que

foram corretamente classificados em relação ao total de casos. Recall positivo é a porcentagem de todos casos de fato positivos que foram classificados como positivo. Recall negativo é a porcentagem de todos casos de fato negativos que foram classificados como negativo [11].

2.4 Processamento da Linguagem Natural

Processamento da Linguagem Natural (PLN) é a fase da mineração de texto em que métodos de pré-processamento são aplicados [9]. O PLN manipula o conteúdo do texto com o objetivo de extrair entidades, detectar relacionamentos e sinônimos e desambiguar o texto através de técnicas como remoção de *stop words*, *tokenization*, *Lemmatization (Part of Speech)* e *steming* [9]. Neste trabalho a ordem dos métodos de pré-processamento aplicados nos textos é *tokenization* seguido de *Lemmatization* e por último remoção de *stop words*.

a) *Tokenization*

Antes de qualquer processamento de texto ser feito, é necessário o texto ser segmentado em unidades linguísticas como palavras, sentenças ou frases [9]. Esse processo é chamado *Tokenization*. A língua adotada por esse trabalho para ser usada na etapa de PLN foi o Inglês. Normalmente as palavras são separadas por espaços em branco, mas existem exceções como o caso do apóstrofo em Inglês (“*I’m*” em que a apóstrofe junta as palavras “*I*” e “*am*”).

Para esse projeto foi adotado o protocolo Penn Treebank[12] para a fase de *Tokenization*. Esse protocolo é um conjunto de regras para a classificação e divisão de textos em Inglês. Figura 3 mostra um exemplo de classificação e divisão de texto feito de acordo com protocolo.

Original :
I want someone to look at me the way Andre De Grasse looks at Usain Bolt!

Classificação feita de acordo com Protocolo Penn Treebank:
I/PP want/VBP someone/NN to/TO look/VB at/IN me/PP the/DT way/NN Andre/NP De/NP Grasse/NP looks/VBZ at/IN Usain/NP Bolt/NP !/.

Depois da Tokenization :
[!] [want] [someone] [to] [look] [at] [me] [the] [way] [Andre] [De] [Grasse] [looks] [at] [Usain] [Bolt] [!]

Figura 3 – Exemplo do protocolo Penn Treebank e do Tokenization. Fonte: SANTORINI, Beatrice [12].

b) Lemmatization e Stemming

O objetivo de *lemmatization* e *stemming* é reduzir as formas variantes de uma mesma palavra [9]. *Lemmatization* vem da palavra *lemma*, que na área de Linguística se refere à forma canônica ou forma do dicionário de uma palavra. O método *Lemmatization* considera o contexto que a palavra está inserida utilizando o *Part-of-speech* para análise morfológica do texto [13]. Com o resultado da análise morfológica do texto o algoritmo identifica a palavra relacionando-a com sua forma canônica, que está no dicionário. O algoritmo de *lemmatization* trabalha junto com um repositório que possui todas as palavras em forma canônica [13]. O método *Stemming* foca em reduzir palavras por um processo de redução de sufixos e prefixos para se encontrar o *stem* (em português radical) sem considerar o contexto que a palavra está inserida. Um exemplo das diferenças dos métodos é quando apresentado o token [saw]. O método de *stemming* provavelmente retorna apenas a letra [s], enquanto *lemmatization* irá tentar retornar [see] ou [saw] dependendo se o *token* está sendo usado como verbo ou substantivo [9]. *Lemmatization* e *Stemming* são métodos populares de pré-processamento e opções excludentes. *Lemmatization* foi escolhido para esse trabalho e o *Part-of-speech* utilizado para apoiar a fase de *Lemmatization* foi o Stanford POS Tagger[13] que utiliza o mesmo protocolo Penn Treebank da fase de *tokenization*.

c) Remoção de *stop words*

Stop words, ou palavras vazias em português, são palavras encontradas com grande frequência em textos e sem significado semântico relevante [9]. Classes gramaticais como pronomes, artigos, numerais, preposição, conjunção e interjeição (e no caso da língua inglesa alguns verbos auxiliares) são considerados *stop words*. No contexto de uma classificação em que se deseja classificar o texto, as *stop words* podem ser erroneamente aprendidas pelos algoritmos como palavras com grande relação a determinada classe quando na verdade são palavras que estão relacionadas a todas as classes e não a uma classe específica [9]. As *stop words* têm função em um texto de melhorar sua compreensão na leitura humana, mas não para diferenciá-lo de outros textos para os algoritmos. Nesse trabalho foi adotada a lista de *stop words* fornecida pelo projeto Stanford CoreNLP.[5]

2.5 Named Entity Recognizer

Named Entity Recognizer (NER), em português reconhecimento de entidades nomeadas, é um classificador de sequências de palavras em textos que podem ser nomes de pessoas, empresas ou locais [14].

O algoritmo NER utilizado neste trabalho é o Stanford NER[14] que é um *CRF Classifier* que, a partir do treinamento de um modelo é capaz de extrair entidades de textos. O Stanford NER utiliza os *tokens* que foram gerados no pré-processamento pela etapa de *tokenization* no protocolo Penn Treebank[12] e classificados como substantivos próprios. Os *tokens* classificados como substantivos próprios são identificados pelo algoritmo em nomes de pessoas, nomes de lugares ou nomes de entidades [14].

2.6 Análise de sentimentos em textos

Análise de sentimento é a tarefa de identificar pontos de vista sobre determinado assunto [15]. A análise de sentimentos em textos é uma atividade de classificação de textos conforme a polaridade das opiniões contidas nos textos [16]. Nos últimos anos muitas técnicas de análise de sentimentos surgiram e cada uma delas tem vantagens e desvantagens em relação a possíveis formatos de textos [6]. De acordo com Ribeiro *et al.* [6] no contexto de textos retirados de redes sociais e classificados em 3 classes (classes positivo, negativo e neutro), a melhor técnica de classificação é o Umigon [17]. Umigon é um projeto *open source* de um classificador de sentimentos baseado em características léxicas de um texto, com foco em textos da rede social Twitter [17]. De acordo com Levallois [17], a classificação é feita em 4 partes: (a) Decomposição do *tweet*, (b) Detecção semântica, (c) Avaliação de *hashtag*, (d) Aplicação de heurísticas.

- a) Decomposição do *tweet*: Decomposição do *tweet* em n-grams com comparação com os termos léxicos para cada n-gram.
- b) Detecção semântica: Detecção semântica do *tweet* com análise de *emojis* baseados em textos e onomatopeias.
- c) Avaliação de *hashtag*: Avaliação de *hashtag* no *tweet*.
- d) Aplicação de heurísticas: Heurística para classificação final do *tweet* em positivo, negativo ou neutro.

O algoritmo divide o *tweet* em n-grams e tenta classificar cada conjunto com base na comparação dos termos com o conjunto de termos já conhecidos. Um termo já conhecido é um termo que tem um sentimento já associado a ele ou possui uma árvore de decisão associada ao termo que pode levar a diferentes classificações [17]. Todos os subconjuntos do *tweet* são classificados com sentimentos diferentes para um mesmo *tweet*. Com todos os n-grams classificados, heurísticas de resolução de conflito são aplicadas. Essas heurísticas identificam a presença de termos moderadores, como por

exemplo os termos “mas”, “apesar” ou “embora” [17]. As heurísticas também identificam termos intensificadores de sentimento que concedem um peso maior a classificação individual de um subconjunto, como por exemplo “muito”, “bastante” e palavras de baixo calão. Algumas negações que ainda não foram computadas na fase de classificação individual do n-gran também são computadas na fase de heurísticas. Nas heurísticas também são consideradas a presença de *hashtags* com comparação de cada parte da *hashtag* com os termos léxicos [17]. Onomatopeias e *emojis* também são identificados e recebem um sentimento associado. A resolução de conflito considera a presença de sentimentos negativos, sentimentos positivos, presença de termos moderados, negações, *emojis*, onomatopeias e termos intensificadores de sentimento para classificar o *tweet* em um sentimento. Caso não haja nenhum sentimento marcado ou que a heurística não consiga distinguir da marcação uma clara evidência de positivo ou negativo o texto então é classificado como neutro.

2. 7 Redes de coocorrência entre palavras

Redes de coocorrência entre palavras são diagramas capazes de mostrar estruturas e padrões de coocorrência entre palavras em um texto. De acordo com Higuchi [18], as palavras são ligadas com base na frequência com que aparecem próximas em textos. O algoritmo considera fatores de frequência da palavra e coocorrência da palavra com outras palavras dada pelo coeficiente jacquard, para decidir quais palavras são expostas no diagrama e quais não são.

Para a elaboração deste diagrama, considera-se o grau de centralidade de cada palavra. O grau de centralidade reflete a proximidade de uma palavra com pelo menos 2 palavras adjacentes que estão no diagrama [18]. O grau de centralidade é maior com base na maior frequência que a proximidade com as duas palavras adjacentes acontece. Linhas são a representação de coocorrência entre os nós, que são as palavras.

Neste trabalho foi imposto um limite para 14 linhas por diagrama para facilitar a

compreensão do diagrama, e foram adotados padrões de cores da tonalidade azul para a tonalidade branca e depois para a rosa para representar o grau de centralidade das palavras nos textos, onde rosa são palavras mais centrais, azuis menos centrais e brancas variantes entre azul e rosa. Neste trabalho foi utilizado o método de diagrama feito por Fruchterman & Reingold[19], em conjunto com as técnicas pré-processamento *Part-of-Speech*, remoção de *stop words*, remoção de caracteres especiais e transformação de todas as letras de maiúsculas para minúsculas. Algumas *stop words* extras foram adicionadas por serem comuns ao tema e não acrescentarem no diagrama como “Rio de Janeiro”, “*athlete*”, “*match*”, “*olympic*”, “*man*”, “*woman*” e “*team*”.

3 Processo de descoberta de desejos

O processo de descoberta de desejos é o conjunto de técnicas e métodos aplicados para encontrar desejos em textos. Esse processo inclui a aplicação dos algoritmos de pré-processamento, de operações de núcleo de descoberta de padrões e de pós-processamento. O domínio escolhido para este trabalho foram os Jogos Olímpicos do Rio de Janeiro de 2016 e mais especificamente os desejos relacionados aos atletas dos jogos.

3.1 Visão geral

O processo começa quando os textos em que se deseja realizar descoberta de desejos estão disponíveis. O processo utiliza técnicas NLP, mineração de texto, NER, análise de sentimentos e redes de coocorrência na ordem que são mencionadas. A Figura 4 mostra o processo de descoberta de desejos.

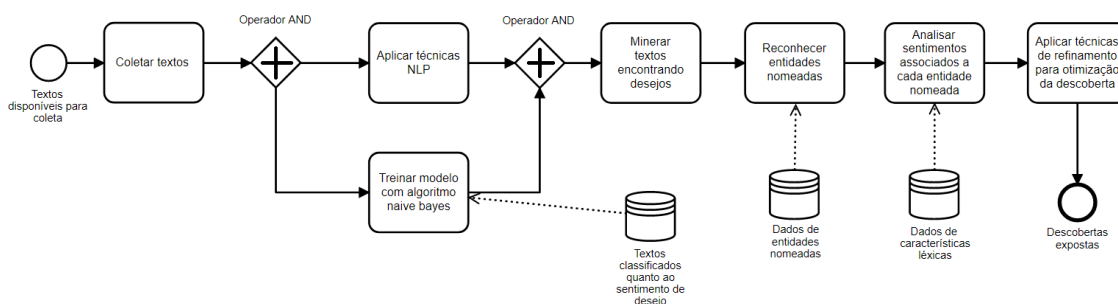


Figura 4 – Processo de descoberta de desejos. Fonte: Autor

3.2 Coleta

A coleta começou meia noite do dia 5 de agosto de 2016, dia da cerimônia de abertura dos jogos olímpicos, até 23:59 do dia 21 de agosto de 2016, dia da cerimônia de encerramento dos jogos olímpicos, totalizando 17 dias corridos de captura. Foi utilizada a biblioteca em Java Twitter4J para conexão com a API do Twitter[20] responsável por fornecer em streaming todas os *tweets*, *retweets* e *replies* da *hashtag* “#Rio2016”, com caracteres independentes de maiúsculo e minúsculo. A *hashtag* #Rio2016 foi escolhida por ser uma *hashtag* usada por usuários de vários idiomas diferentes e por ter sido a *hashtag* adotada oficialmente pelo Twitter para o evento [7].

Os dados foram capturados no formato:

- uma *string* para o texto do *tweet/retweet/reply*.
- uma *string* para o nome do usuário.
- uma *string* para data e hora da criação do texto.
- uma variável booleana que indica se é *tweet* ou *retweet*.
- uma *string* para o idioma do texto.
- uma *string nullable* para nomes de usuários que se estão fazendo um *reply*.

Inicialmente os dados foram salvos em formato Excel (xlsx), para facilitar seleção e limpeza de dados, e então exportados para formato de texto (txt). Ao longo dos 17 dias 4,351,740 de *tweets* em português foram capturados na #Rio2016 e 12,010,517 de *tweets* em inglês na mesma *hashtag*.

Tabela 1 - Quantidade de tweets, retweets e replies coletadas durante etapa de coleta para língua portuguesa.

Fonte : Autor

Dia	Total Bruto	Número de retweets	Número de tweets mais replies
5	503197	327519	175678
6	479533	311739	167794
7	290074	191638	98436
8	203009	119734	83275
9	180304	100761	79543
10	175727	85791	89936
11	191943	110089	81854
12	104620	57630	46990
13	290907	235279	55628
14	201765	122180	79585
15	126259	78964	47295
16	279396	170898	108498
17	162702	102072	60630
18	172453	98060	74393
19	100911	65380	35531
20	388113	252672	135441
21	500827	344684	156143

Tabela 2 - Quantidade de tweets, retweets e replies coletadas durante etapa de coleta para língua inglesa.

Fonte : Autor

Dia	Total Bruto	Número de retweets	Número de tweets mais replies
5	866201	642734	223467
6	600611	436237	164374
7	577694	401447	176247
8	628518	460045	168473
9	601160	436782	164378
10	603464	398004	205460
11	919081	615116	303965
12	619004	413839	205165
13	589610	405828	183782
14	999396	658815	340581
15	620082	448393	171689
16	660188	453564	206624
17	528654	374262	154392
18	902051	644544	257507
19	715869	499796	216073
20	874266	657345	216921
21	704668	486789	217879

3.3 Tradução

Neste trabalho na etapa de captura foram obtidos textos em duas línguas diferentes o Português e o Inglês, mas para a parte de processamento foi escolhido unificar a língua para poder utilizar as mesmas análises em todos os textos e assim poder comparar resultados. Como a maior parte dos textos obtido estão na língua inglesa foi escolhido traduzir todos os textos em Português para o Inglês.

Foi utilizada a tradução fornecida pelo Google Tradutor [21], através de uma função na ferramenta Google Sheets. Vale ressaltar que a ferramenta Google Sheets só permite traduzir de 2000 em 2000 textos, cada texto contendo até 5000 caracteres. Como um *tweet* tem no máximo 140 caracteres o limite de caracteres por texto não foi atingido em nenhum momento. Para a tradução massiva dos textos de português para o inglês, foi utilizada a linguagem de automação AutoIt [22]. AutoIt conta com uma integração com o Microsoft Excel e o Internet Explorer, de forma que foi possível copiar de 2000 em 2000 campos do Excel para o Explorer, realizar a tradução no Google Sheets e retornar com a tradução feita para o Excel.

3.4 Pré-processamento - técnicas PLN

Considerando a particularidade da rede social Twitter e as diferenças entre as formas de se expressar em *tweets*, *retweets* e *replies* foram feitas seleções. Uma seleção levou em conta o esforço para um usuário de se comunicar utilizando cada uma das formas de se expressar. Enquanto no *tweet* e *reply* o usuário é forçado a escrever um texto, no *retweet* escrever um texto é opcional. Por se tratar de uma mineração de textos assumiu-se que para a opinião do usuário ser considerada ele deveria ter que passar pelo esforço de escrever um texto. Logo, os *retweets* foram eliminados. Os *replies* foram

mantidos pela interessante construção de um diálogo entre opiniões e sua capacidade de enriquecer a descoberta de desejos.

Os *tweets* e *replies* foram selecionados para o pré-processamento. Após o uso dos campos de data e hora de criação do texto, e idioma do texto para organização das tabelas em formato “.xlsx”, o texto em si foi exportado para arquivos no formato “.txt”. Os textos foram colocados no formato do “.txt” em “UTF-8” para a unificação de formatação.

3.5 Treinamento Naive Bayes para identificação de desejos

O treinamento do Naive Bayes foi feito utilizando a junção de dois corpus, e com validação cruzada (10 *folds cross-validation*). O primeiro corpus continha os *reviews* de produtos extraído do amazon.com e cnet.com de Hu e Liu [23]. O segundo corpus continha os posts de discussão política extraído do politics.com por Mullen e Malouf [23]. Os corpus foram escolhidos por apresentar domínios diferentes o que mostra que o modelo não será construído para identificação de desejos em apenas um domínio específico.

O modelo foi treinado e conseguiu uma acurácia inicial de 81%. Para melhorar o modelo, foram utilizados para o treinamento filtros baseados no *Part-of-speech*, removendo palavras marcadas como substantivos, substantivos próprios, expressões estrangeiras, preposições, ficando apenas com palavras marcadas como adjetivos, verbos e advérbios para treinamento, com isso, a acurácia aumentou para 85%. Vale ressaltar que esse modelo com acurácia de 85% apresentou número de falsos positivos baixo com precisão positiva de 97%, precisão negativa de 80%, mas a recall positivo também ficou baixa em 64% e recall negativo de 98%. Analisando uma amostra de *tweets* classificados percebeu-se que o modelo ficou atrelado a verbos relacionados ao sentimento de desejo. Isso gerou um modelo de alta precisão com poucos falso-positivos, mas com grande quantidade de falso-negativos. Por esta etapa do trabalho

poder contar com uma base já selecionada de 1,576,650 de *tweets* em português e 3,576,977 de *tweets* em inglês mesmo com grande quantidade de falso-negativos ainda foi possível extrair uma grande quantidade de textos com desejos. A alta precisão gerada por poucos falso-positivos permitiu que os resultados dos textos classificados como positivos fosse uma representação adequada de textos com o sentimento de desejo, o que era o objetivo desta etapa do processo.

3.6 Reconhecimento de Entidades Nomeadas - NER

O Software Stanford NER [14] foi utilizado para o reconhecimento de entidades nomeadas nos textos classificados como desejo. O NER utiliza um modelo de treinamento para aprender a reconhecer e distinguir as diferentes classes de entidade. O modelo de treinamento utilizado neste trabalho foi o disponibilizado no CoNLL-2003 base *eng.train* [14].

Este modelo é capaz de distinguir três classes de entidade: nome de pessoas, nome de locais e nome de organizações. O caractere referente a *hashtag* “#” foi retirado dos textos para que nomes dentro de *hashtags* fossem considerados. Como o objetivo do trabalho é identificar desejos para os atletas dos jogos olímpicos a única classe de entidade nomeada utilizada foi o nome de pessoas.

No final desta etapa, foi possível identificar 5,182,609 citações a entidades nomeadas nos textos em Inglês e 1,939,480 nos textos em Português. A Figura 5 ilustra um extrato da saída do Stanford NER, com alguns dos nomes de atletas reconhecidos.

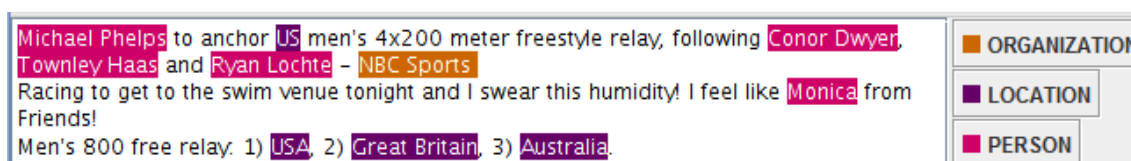


Figura 5 – Screenshot da GUI do Stanford NER[5]. Fonte: Autor

3.7 Análise de sentimentos em textos

Com a lista de nomes de pessoas encontrado pelo NER foi possível separar todos os *tweets* referentes a cada pessoa específica. Esses *tweets* foram então classificados em classes: positivo, negativo ou neutro. Com o Umigon [17] foi possível perceber a diferença entre a quantidade de desejos associados a cada atleta em cada etapa dos Jogos Olímpicos, e a polarização da maioria dos desejos. Foi possível também classificar todos os *tweets* para concluir se o sentimento de desejo é mais propenso a ser positivo, negativo ou neutro. Vale ressaltar que o Umigon é mais focado na correção da classificação quanto aos sentimentos polarizados do que quanto a cobertura da classificação, de forma que caso haja uma complexa junção de diferentes sentimentos na mesma frase o algoritmo tende a classificar como neutro do que assumir que a frase é polarizada.

No final desta etapa, foi possível polarizar os *tweets* para de cada entidade nomeada. A Figura 6 ilustra um extrato da saída do Umigon para a entidade nomeada Usain Bolt. Os resultados completos das análises estão presentes em visões no próximo capítulo.

	A	B
1	I honestly got chills witnessing this his greatness last night. He just wins because he wants to. #Rio2016 #Bolt	Positivo
2	Is getting frustrated with BBC coverage when I don't have a red button. I want to see Brits performances and not 2 hours of Bolt #Rio2016	Negativo
3	Wish bolt had 10 brothers and sisters.... #Rio2016	Neutro

Figura 6 – Screenshot da saída do Umigon[17]. Fonte: Autor

4 Análise de descobertas

Este capítulo tem o objetivo de levantar, mapear e relacionar potenciais motivos de desejos.

4.1 Visão de Popularidade Diária

Com base no resultado da descoberta de nomes de entidades aplicado nos *tweets* marcados como desejo, foi possível gerar a Tabela 3, para as 5 pessoas mais comentadas por dia (sendo *Rank 1* a mais comentada, *Rank 2* a segunda mais comentada, e assim por diante) da olimpíada em inglês, e a Tabela 4 para mesmo resultado em português.

Vale ressaltar que alguns nomes (como, por exemplo da Gisele Bündchen) foram escritos de muitas formas diferentes, e foi necessária uma interpretação para unificar as diferentes formas em apenas uma. Donald Trump e Michel Temer também foram muito referenciados apenas pelo sobrenome e também ocorreu uma interpretação para unificar a forma completa do nome com a referência baseada apenas no sobrenome.

Tabela 3 – Rank com quantidade de tweets da língua inglesa marcados como desejo para os 5 atletas mais citados por dia de competição. Fonte: Autor

Dia	Rank 1	Qnt	Rank 2	Qnt	Rank 3	Qnt	Rank 4	Qnt	Rank 5	Qnt
5	Michael Phelps	869	Andy Murray	728	Gisele Bündchen	653	Donald Trump	632	Kevin Durant	597
6	Samir ait adam	715	Michael Phelps	541	Kevin Durant	433	Mack Horton	422	Greg van Avermaet	409
7	Michael Phelps	755	Serena williams	691	Katie Ledecky	524	Nathan Adrian	501	Adam Peaty	493

8	Michael Phelps	663	Serena williams	531	Katie Ledecky	477	Lilly King	452	Ryan Murphy	436
9	Michael Phelps	647	Simone Biles	519	Serena williams	495	Aly Raisman	446	Gabby Douglas	438
10	Katie Ledecky	694	Michael Phelps	638	Allison Schmitt	547	Simone Biles	540	Aly Raisman	534
11	Michael Phelps	859	Simone Biles	747	Simone manuel	633	Ryan Murphy	512	Kayla Harrison	482
12	Katie ledecky	784	Michael phelps	755	Monica Puig	692	Joseph Schooling	547	Maya DiRado	529
13	Michael Phelps	712	Monica Puig	682	Mo Farah	659	Andy Murray	649	Joseph Schooling	612
14	Usain Bolt	1145	Andy Murray	814	Simone Biles	788	Juan Martín del Potro	671	Michael phelps	655
15	Usain Bolt	538	Ryan Lochte	527	Mark Cavendish	511	Jimmy Feigen	493	Jack Conger	490
16	Simone Biles	1246	Laura trott	684	Usain Bolt	672	Jason Kenny	560	Ryan Lochte	543
17	Usain Bolt	617	Andre de grasse	549	Justin Gatlin	510	Simone Biles	491	Tianna Bartoletta	487
18	Usain bolt	643	Andre de grasse	522	Ryan lochte	488	Helen Maroulis	461	Kerron Clement	455
19	Usain bolt	616	Tianna Bartoletta	561	Ryan lochte	524	Lutalo Muhammad	479	Kevin Durant	470
20	Kevin Durant	657	Breanna Stewart	624	Mo Farah	620	Usain bolt	594	Ryan lochte	557
21	Michael Phelps	743	Kevin Durant	662	Breanna Stewart	654	Ryan lochte	643	Usain bolt	635

Tabela 4 – Rank com quantidade de tweets da língua portuguesa marcados como desejo para os 5 atletas mais citados por dia de competição. Fonte: Autor

Dia	Rank 1	Qnt	Rank 2	Qnt	Rank 3	Qnt	Rank 4	Qnt	Rank 5	Qnt
5	Michel Temer	469	Anitta	415	Vanderlei Cordeiro	336	Gisele Bündchen	328	Michael Phelps	320
6	Marta Vieira da Silva	390	Anitta	388	Michael Phelps	372	Felipe Wu	364	Neymar	359
7	Michael Phelps	379	Neymar	364	Annemiek van Vleuten	358	Marta Vieira da Silva	319	Érica Miranda	314
8	Rafaela Silva	462	Michael Phelps	367	Neymar	340	Érica Miranda	292	Diego hypolito	280
9	Michael Phelps	320	Rafaela Silva	317	Neymar	278	Daniele Hypolito	270	Marta Vieira da Silva	238
10	Neymar	330	Michael phelps	312	Rafaela Silva	309	Gabriel Barbosa	240	Galvão Bueno	238
11	Michael Phelps	301	Mayra Aguiar	248	Thiago Pereira	240	Jade Barbosa	232	Neymar	229
12	Marta Vieira da Silva	267	Bárbara Barbosa	248	Rafael Silva	234	Michael Phelps	212	Neymar	204
13	Neymar	240	Michael Phelps	237	Rafaela Silva	221	Usain Bolt	217	Thiago Pereira	210
14	Usain Bolt	334	Neymar	317	Diego Hypólito	288	Michael Phelps	274	Arthur Nory	227
15	Thiago Braz	374	Usain Bolt	364	Neymar	359	Ryan Lochte	347	Arthur Zanetti	276
16	Robson conceicao	345	Isaquias Queiroz	340	Ryan Lochte	317	Thiago Braz	314	Usain Bolt	312

17	Neymar	330	Usain Bolt	314	Robson conceicao	301	Andre de grasse	295	Ágatha Bednarczuk	265
18	Usain bolt	317	Isaquias queiroz	310	Martine grael	304	Neymar	295	Ryan Lochte	287
19	Usain bolt	324	Wallace de Souza	320	Sérgio Santos	293	Ryan Lochte	384	Bruno schmidt	272
20	Neymar	473	Usain bolt	327	Maicon Siqueira	304	Weverton Pereira da Silva	300	Ryan Lochte	276
21	Neymar	501	Wallace de Souza	453	Sérgio Santos	330	Ricardo Lucarelli	315	Galvão Bueno	304

4.2 Visão da Popularidade Condensada

Somando todos os dados da popularidade diária, foi possível criar as Tabela 5 e 6, para os 5 atletas mais “populares” (com mais *tweets* classificados como desejos) durante todos os dias de competição (em Inglês e em Português).

Tabela 5 – Rank com quantidade de tweets da língua inglesa marcados como desejo para os 5 atletas mais citados nos Jogos Olímpicos. Fonte: Autor

Posição Geral	Nome do atleta citado mais vezes em <i>tweets</i> marcados como desejo	Quantidade Geral de <i>tweets</i> em todos os 17 dias
1°	Michael Phelps	10254
2°	Usain Bolt	9029
3°	Simone Biles	8487
4°	Ryan Lochte	7601
5°	Katie Ledecky	7011

Tabela 6 – Rank com quantidade de tweets da língua portuguesa marcados como desejo para os 5 atletas mais citados nos Jogos Olímpicos. Fonte: Autor

Posição Geral	Nome do atleta citado mais vezes em <i>tweets</i> marcados como desejo	Quantidade Geral de <i>tweets</i> em todos os 17 dias
1°	Neymar	5496
2°	Michael Phelps	4622
3°	Usain Bolt	4259
4°	Marta Vieira da Silva	3890
5°	Rafaela Silva	3195

4.3 Visão Desejo

Utilizando os *tweets* marcados como desejo na língua inglesa e filtrando usando o *Part-Of-Speech* é possível encontrar os principais verbos presentes nos *tweets* marcados como desejo dos Jogos Olímpicos. Vale ressaltar que verbos e auxiliares presentes nos *stop words* (como “*be*”, “*can*”, “*do*”, “*have*”, “*will*”, “*would*”, “*take*”, “*must*”, “*ought*” e “*should*”) são retirados.

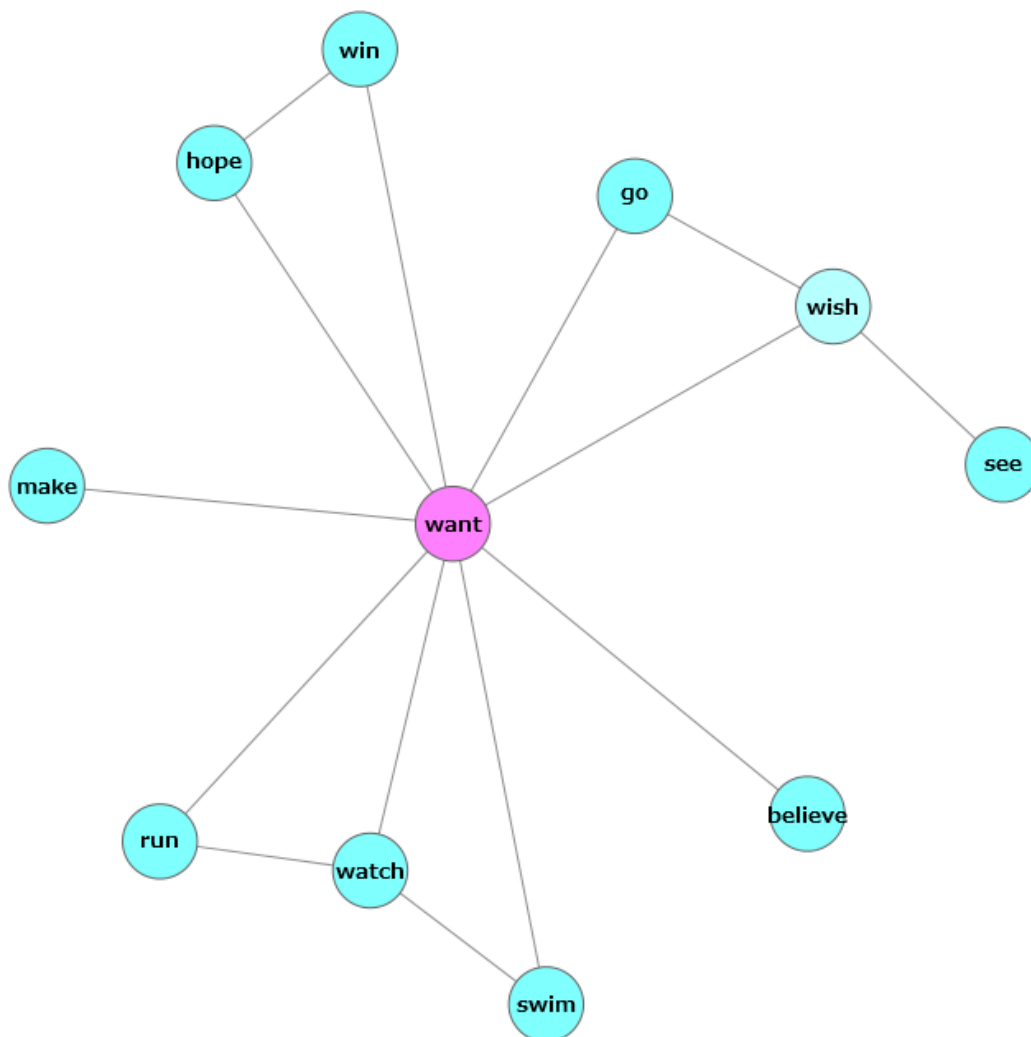


Figura 7 – Diagrama de coocorrência para verbos considerando todos os tweets marcados como desejo da língua inglesa. Fonte: Autor

O verbo com maior frequência nos textos com desejo marcado é o verbo *want*, que significa querer. O querer pode ser considerado parte do sentimento de desejo, de forma que o querer pertence a uma classe maior que seria o desejo. No caso em estudo de pessoas relacionadas aos Jogos Olímpicos o querer relacionado a *win*, que significa vitória (como substantivo ou verbo), obteve o maior coeficiente Jacquard. A grande quantidade da palavra querer relacionado a palavra vitória pode mostra um motivo para desejos.

4.4 Visão Quantitativa dos Sentimentos

Neymar e Michael Phelps foram os atletas alvo de mais desejos considerando todos os dados coletados em Português e Inglês. As Figuras 8 e 9 e as Tabelas 7 e 8 mostram a relação quantitativa de desejos positivos, negativos e neutros. Percebe-se uma regularidade da polarização dos desejos a esses atletas ao longo dos dias da competição. Ainda, a maioria quantitativa de desejos para os dois atletas é neutra seguido por desejos de polarização positiva e a minoria tem polarização negativa.

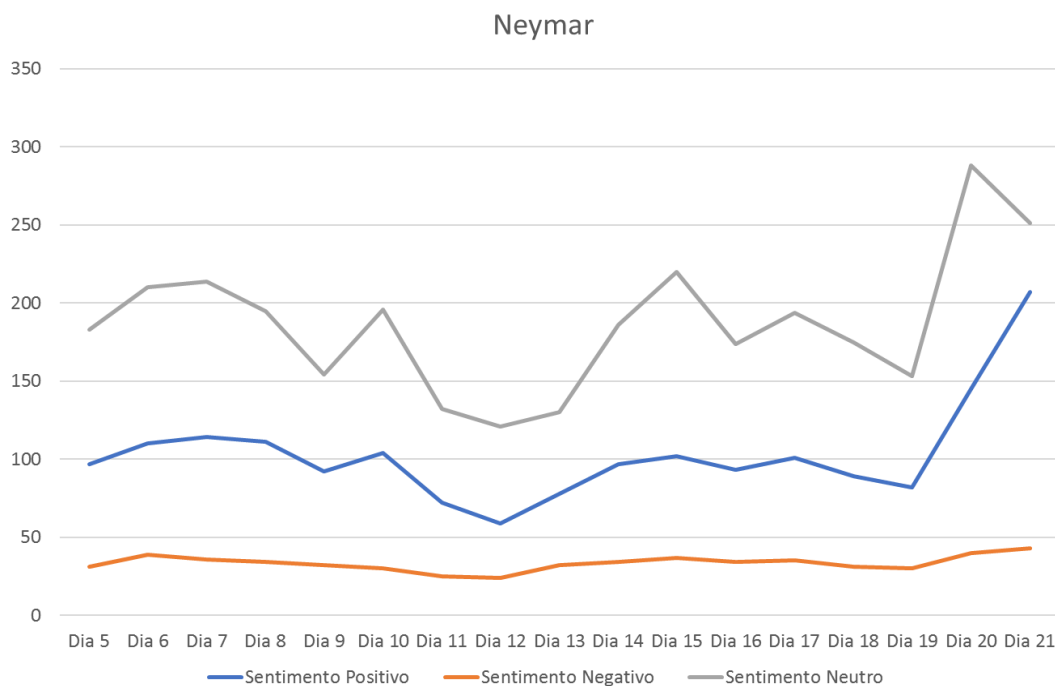


Figura 8 – Gráfico quantitativo da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Neymar. Baseado em tweets da língua portuguesa. Fonte: Autor

Tabela 7 - Tabela quantitativa da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Neymar. Baseado em tweets da língua portuguesa. Fonte: Autor

	Sentimento Positivo	Sentimento Negativo	Sentimento Neutro
Dia 5	97	31	183
Dia 6	110	39	210
Dia 7	114	36	214
Dia 8	111	34	195
Dia 9	92	32	154
Dia 10	104	30	196
Dia 11	72	25	132
Dia 12	59	24	121
Dia 13	78	32	130
Dia 14	97	34	186
Dia 15	102	37	220
Dia 16	93	34	174
Dia 17	101	35	194
Dia 18	89	31	175
Dia 19	82	30	153
Dia 20	145	40	288
Dia 21	207	43	251

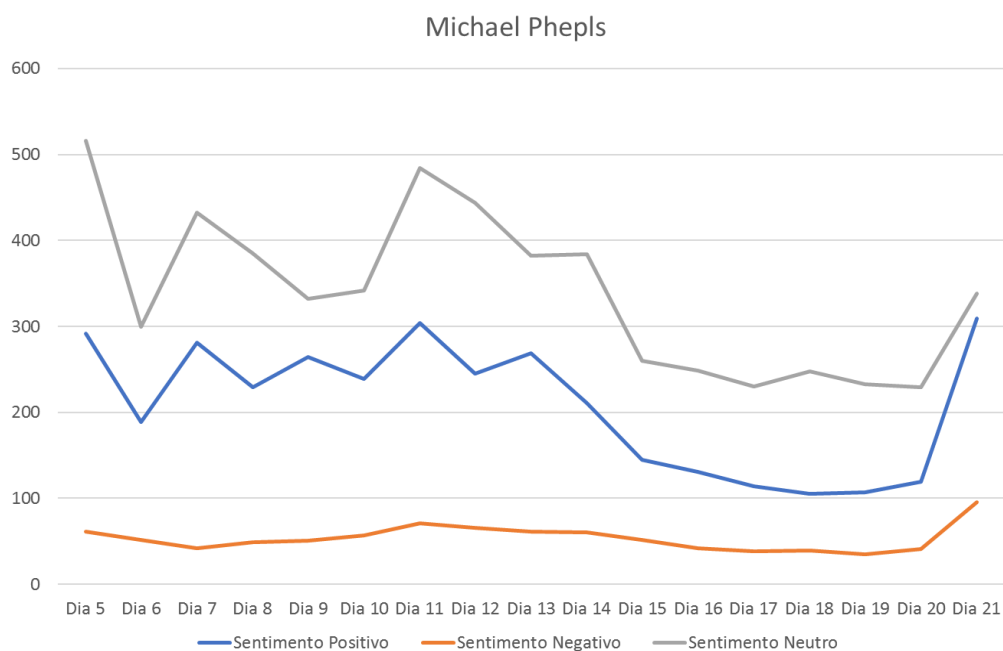


Figura 9 – Gráfico quantitativo da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Michael Phelps. Baseado em tweets da língua inglesa. Fonte: Autor

Tabela 8 - Tabela quantitativa da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Michael Phelps. Baseado em tweets da língua inglesa. Fonte: Autor

	Sentimento Positivo	Sentimento Negativo	Sentimento Neutro
Dia 5	292	61	516
Dia 6	189	52	300
Dia 7	281	42	432
Dia 8	229	49	385
Dia 9	264	51	332
Dia 10	239	57	342
Dia 11	304	71	484
Dia 12	245	66	444
Dia 13	269	61	382
Dia 14	211	60	384
Dia 15	145	52	260
Dia 16	131	42	249
Dia 17	114	38	230
Dia 18	105	39	248
Dia 19	107	35	233
Dia 20	119	41	229
Dia 21	309	96	338

O resultado quantitativo de sentimentos para atletas como Michael Phelps, Usain Bolt e Neymar é praticamente uma constante em relação a *tweets* positivos sempre em valor quantitativo maior que negativos. Para muitos atletas o sentimento em relação a eles não muda durante a olimpíada se mantendo uma constante sem que a relação quantitativa entre positivo, negativo e neutro mude, mas para alguns atletas a relação quantitativa entre sentimentos apresenta mudanças. Para Serena Williams, Ryan Lochte, Diego Hypólito e Marta Vieira da Silva a relação apresenta um comportamento peculiar, por isso foram escolhidos para apresentação.

(A) Serena Williams

A relação quantitativa para Serena Williams, feita com base nos *tweets* na língua inglesa apresentou um comportamento peculiar, pois no dia 09/08/2016 o sentimento negativo pela primeira vez se tornou numericamente maior que o sentimento positivo. Vale ressaltar que nesse mesmo dia a atleta foi eliminada do torneio de tênis da olimpíada.

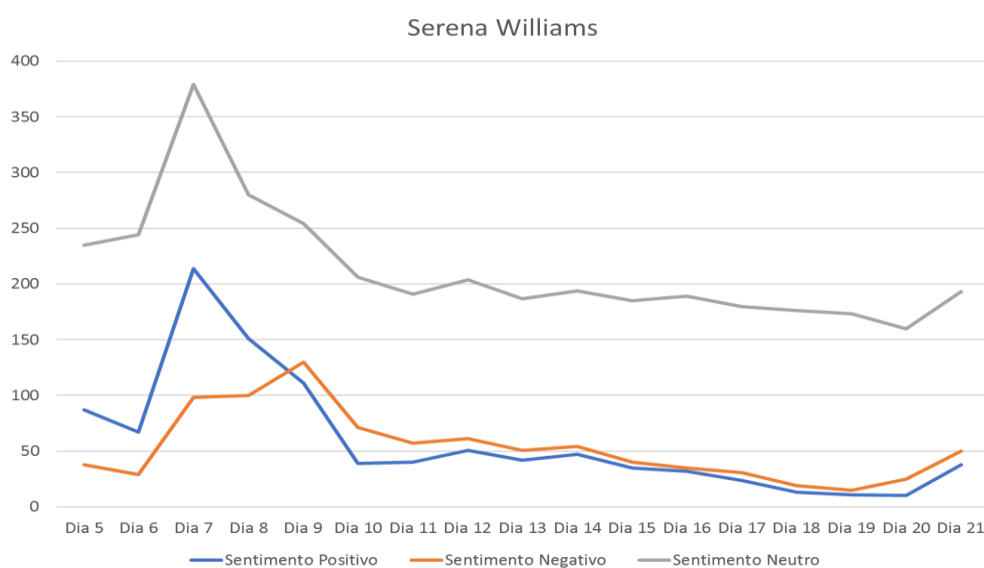


Figura 10 – Gráfico quantitativo da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Serena Williams. Baseado em tweets da língua inglesa. Fonte: Autor

Tabela 9 - Tabela quantitativa da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Serena Williams. Baseado em tweets da língua inglesa. Fonte: Autor

	Sentimento Positivo	Sentimento Negativo	Sentimento Neutro
Dia 5	87	38	235
Dia 6	67	29	244
Dia 7	214	98	379
Dia 8	151	100	280
Dia 9	111	130	254
Dia 10	39	71	206
Dia 11	40	57	191
Dia 12	51	61	204
Dia 13	42	51	187
Dia 14	47	54	194
Dia 15	35	40	185
Dia 16	32	35	189
Dia 17	24	31	180
Dia 18	13	19	176
Dia 19	11	15	173
Dia 20	10	25	160
Dia 21	38	50	193

(B) Ryan Lochte

A relação quantitativa para Ryan Lochte, feita com base nos *tweets* na língua inglesa apresentou um comportamento peculiar, pois no dia 16/08/2016 o sentimento negativo pela primeira vez se tornou numericamente maior que o sentimento positivo. Vale ressaltar que nesse mesmo dia foram exibidos os vídeos que comprovam uma falsa comunicação de um crime de assalto.

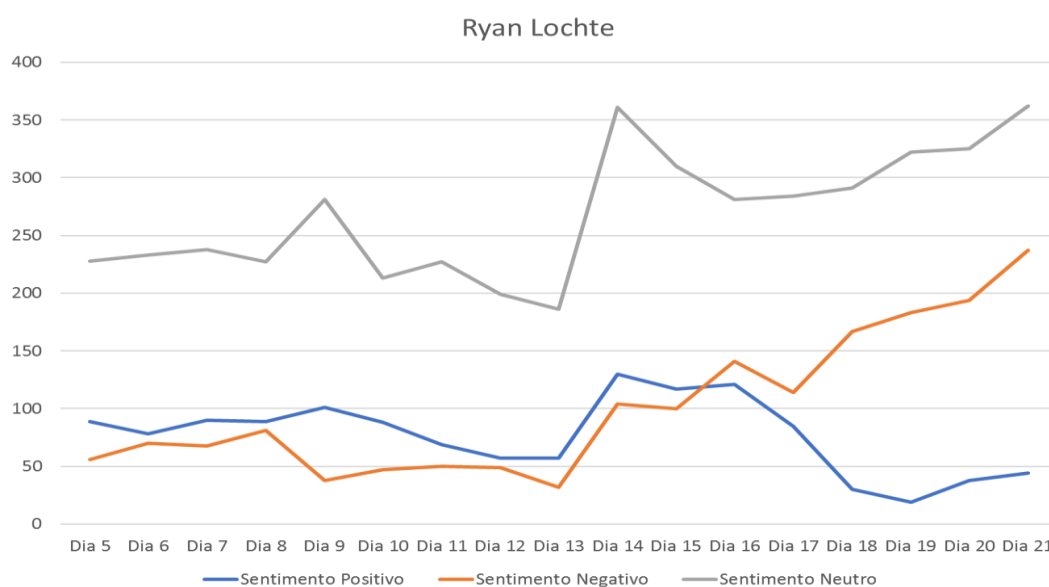


Figura 11 – Gráfico quantitativo da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Ryan Lochte. Baseado em tweets da língua inglesa. Fonte: Autor

Tabela 10 - Tabela quantitativa da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Ryan Lochte. Baseado em tweets da língua inglesa. Fonte: Autor

	Sentimento Positivo	Sentimento Negativo	Sentimento Neutro
Dia 5	89	56	228
Dia 6	78	70	233
Dia 7	90	68	238
Dia 8	89	81	227
Dia 9	101	38	281
Dia 10	88	47	213
Dia 11	69	50	227
Dia 12	57	49	199
Dia 13	57	32	186
Dia 14	130	104	361
Dia 15	117	100	310
Dia 16	121	141	281
Dia 17	85	114	284
Dia 18	30	167	291
Dia 19	19	183	322
Dia 20	38	194	325
Dia 21	44	237	362

(C) Diego Hypólito

A relação quantitativa para Diego Hypólito, feita com base nos *tweets* na língua portuguesa apresentou um comportamento peculiar, pois no dia 14/08/2016 o sentimento positivo pela primeira vez se tornou numericamente maior que o sentimento negativo. Vale ressaltar que no mesmo dia 14/08/2016 o atleta ganhou medalha de prata na competição de ginástica artística.

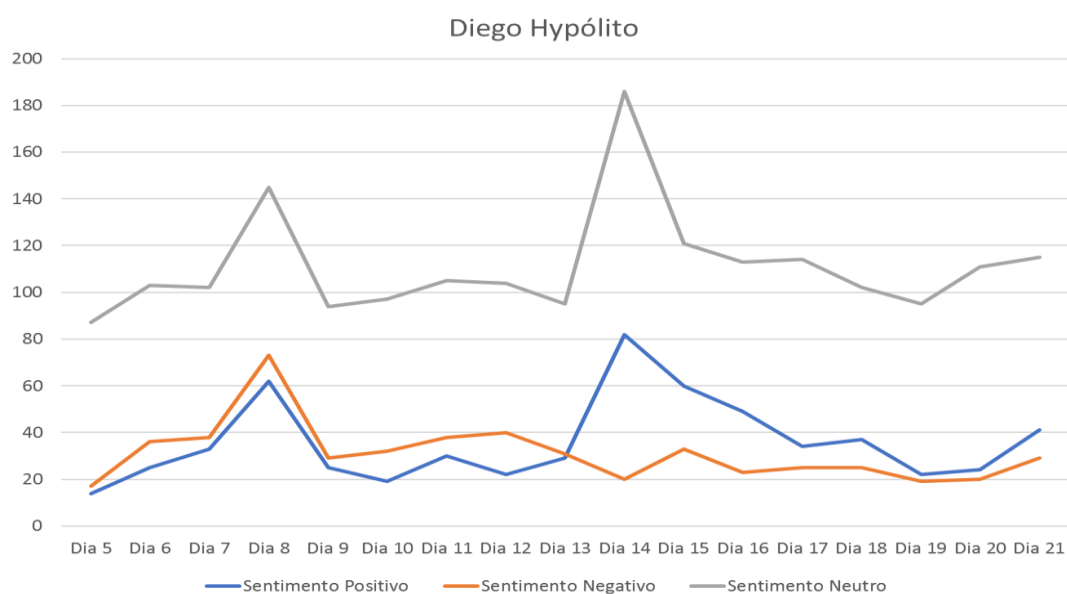


Figura 12 – Gráfico quantitativo da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Diego Hypólito. Baseado em tweets da língua portuguesa. Fonte: Autor

Tabela 11- Tabela quantitativa da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Diego Hypólito. Baseado em tweets da língua portuguesa. Fonte: Autor

	Sentimento Positivo	Sentimento Negativo	Sentimento Neutro
Dia 5	14	17	87
Dia 6	25	36	103
Dia 7	33	38	102
Dia 8	62	73	145
Dia 9	25	29	94
Dia 10	19	32	97
Dia 11	30	38	105
Dia 12	22	40	104
Dia 13	29	31	95
Dia 14	82	20	186
Dia 15	60	33	121
Dia 16	49	23	113
Dia 17	34	25	114
Dia 18	37	25	102
Dia 19	22	19	95
Dia 20	24	20	111
Dia 21	41	29	115

(D) Marta Vieira da Silva

A relação quantitativa para Marta Vieira da Silva, feita com base nos *tweets* na língua portuguesa apresentou um comportamento peculiar, pois no dia 16/08/2016 o sentimento negativo pela primeira vez se tornou numericamente maior que o sentimento positivo. Vale ressaltar que nesse dia a seleção feminina brasileira perdeu para a Suécia.

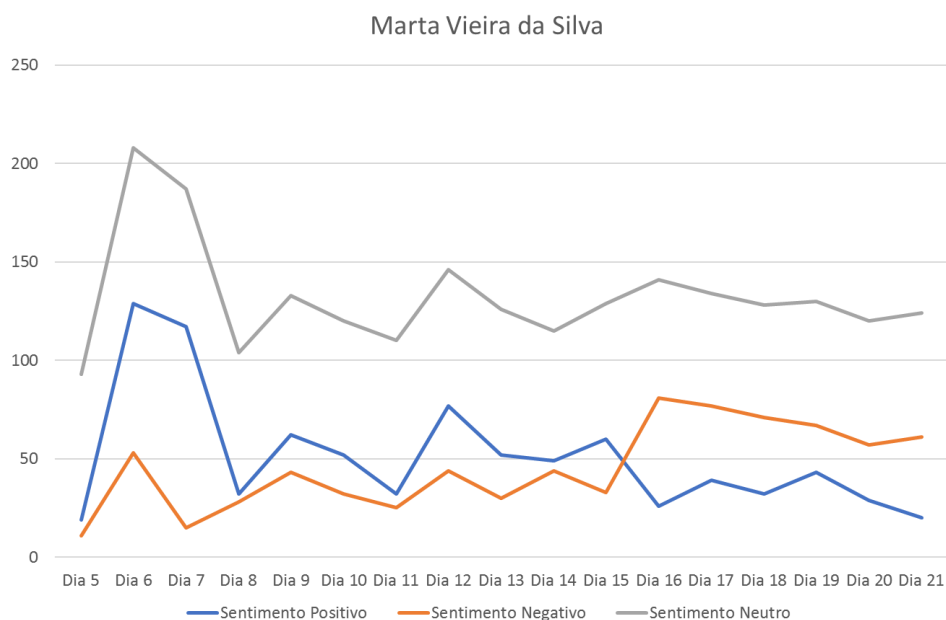


Figura 13 – Gráfico quantitativo da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Marta Vieira da Silva. Baseado em tweets da língua portuguesa. Fonte: Autor

Tabela 12- Tabela quantitativa da classificação positivo, negativo e neutro de tweets marcados como desejo para atleta Marta Vieira da Silva. Baseado em tweets da língua portuguesa. Fonte: Autor

	Sentimento Positivo	Sentimento Negativo	Sentimento Neutro
Dia 5	19	11	93
Dia 6	129	53	208
Dia 7	117	15	187
Dia 8	32	28	104
Dia 9	62	43	133
Dia 10	52	32	120
Dia 11	32	25	110
Dia 12	77	44	146
Dia 13	52	30	126
Dia 14	49	44	115
Dia 15	60	33	129
Dia 16	26	81	141
Dia 17	39	77	134
Dia 18	32	71	128
Dia 19	43	67	130
Dia 20	29	57	120
Dia 21	20	61	124

4.5 Visão de Coocorrência

Redes de coocorrência entre palavras conseguem mostrar a relação entre palavras que levam a um texto ser classificado como desejo positivo ou desejo negativo. Alguns casos como Ryan Lochte e Marta Vieira da Silva mostram possíveis motivos que levam a uma pessoa receber uma grande quantidade de *tweets* com desejo positivo ou negativo.

(a) Ryan Lochte

Do dia 05/08/2016 até 15/08/2016 Ryan Lochte foi citado em 965 *tweets* positivos na língua inglesa. A coocorrência entre as palavras nesses 965 *tweets* positivos, no diagrama da Figura 14, apresenta a palavra *gold*, que significa ouro em Português. A palavra *gold* pode mostrar um motivo de desejos com sentimentos positivos.

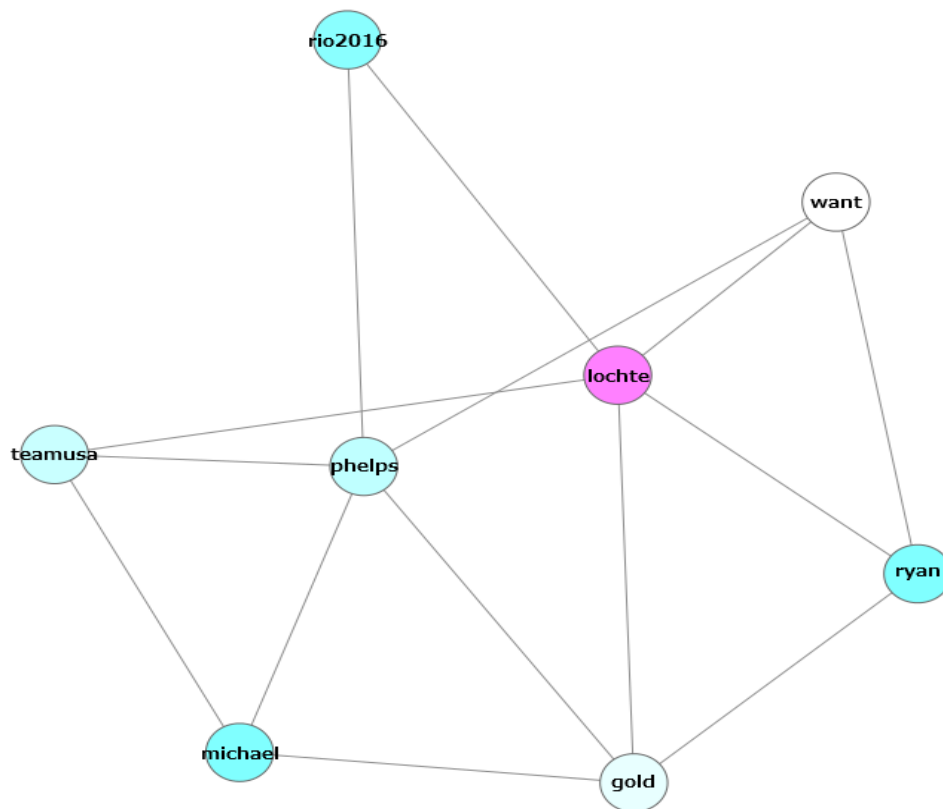


Figura 14 – Diagrama de coocorrência para atleta Ryan Lochte de 05/08/2016 até 15/08/2016. Fonte: Autor

Do dia 16/08/2016 até 21/08/2016 Ryan Lochte foi citado em 1036 *tweets* negativos na língua inglesa. A coocorrência entre as palavras nesses 1036 *tweets* negativos, no diagrama da Figura 15, apresenta a palavra *scandal*, que significa escândalo em Português. A palavra *scandal* pode mostrar um motivo para desejos com sentimento negativo.

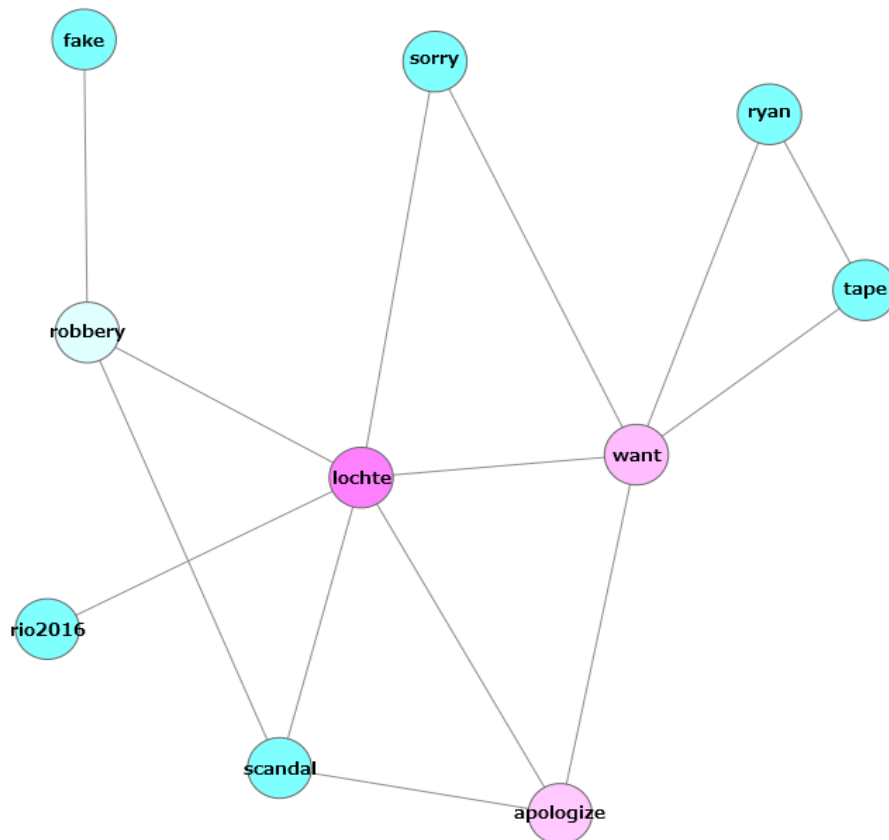


Figura 15 – Diagrama de coocorrência para atleta Ryan Lochte de 16/08/2016 até 21/08/2016. Fonte: Autor

(b) Marta Vieira da Silva

Do dia 05/08/2016 até 15/08/2016 Marta Vieira da Silva foi citado em 681 *tweets* marcados como desejo positivo na língua portuguesa. A coocorrência entre as palavras nesses 681 *tweets* positivos, no diagrama da Figura 16, apresenta a palavra *goal*, que significa objetivo em Português. A palavra *goal* pode mostrar um motivo para desejos com sentimento positivo.

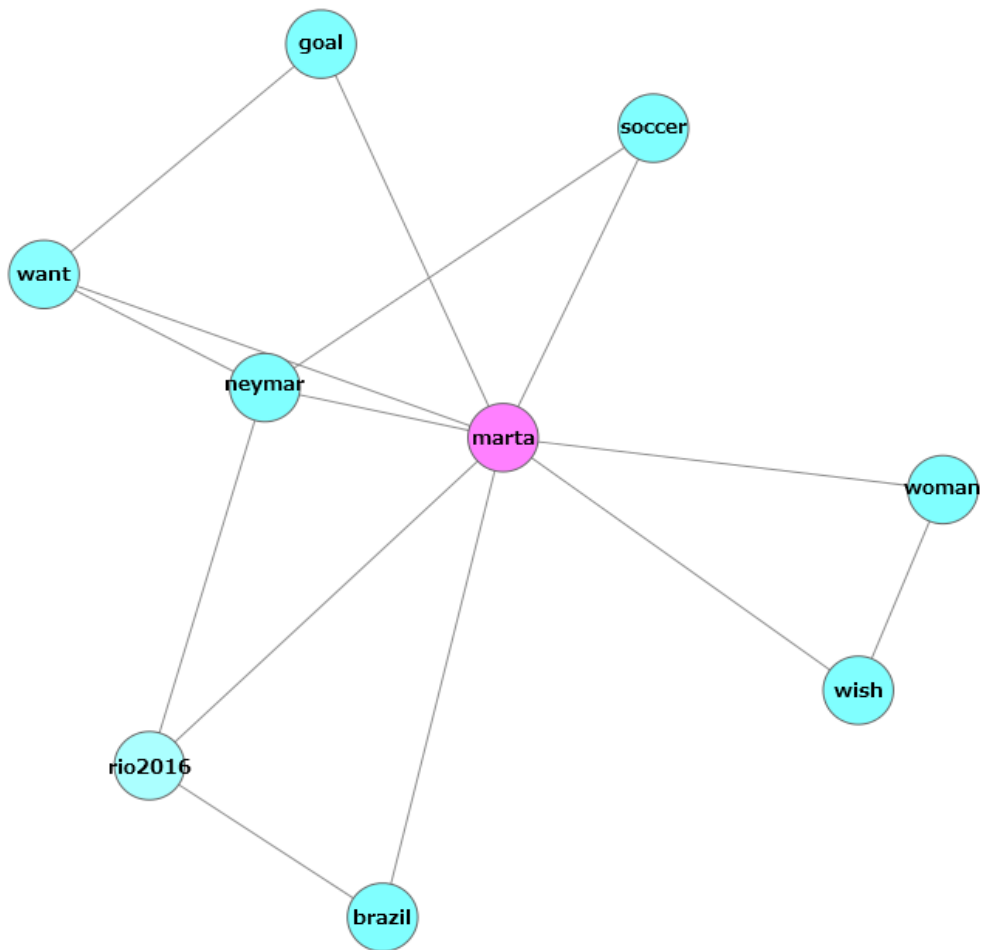


Figura 16 – Diagrama de coocorrência para atleta Marta Vieira da Silva de 05/08/2016 até 15/08/2016. Fonte: Autor

Do dia 16/08/2016 até 21/08/2016 Marta Vieira da Silva foi citada em 414 *tweets* marcados como desejo negativo na língua portuguesa. A coocorrência entre as palavras nesses 414 *tweets* negativos, no diagrama da Figura 17, apresenta a palavra *lose*, que significa perder em Português. A palavra *lose* pode mostrar um motivo para desejos com sentimento negativo.

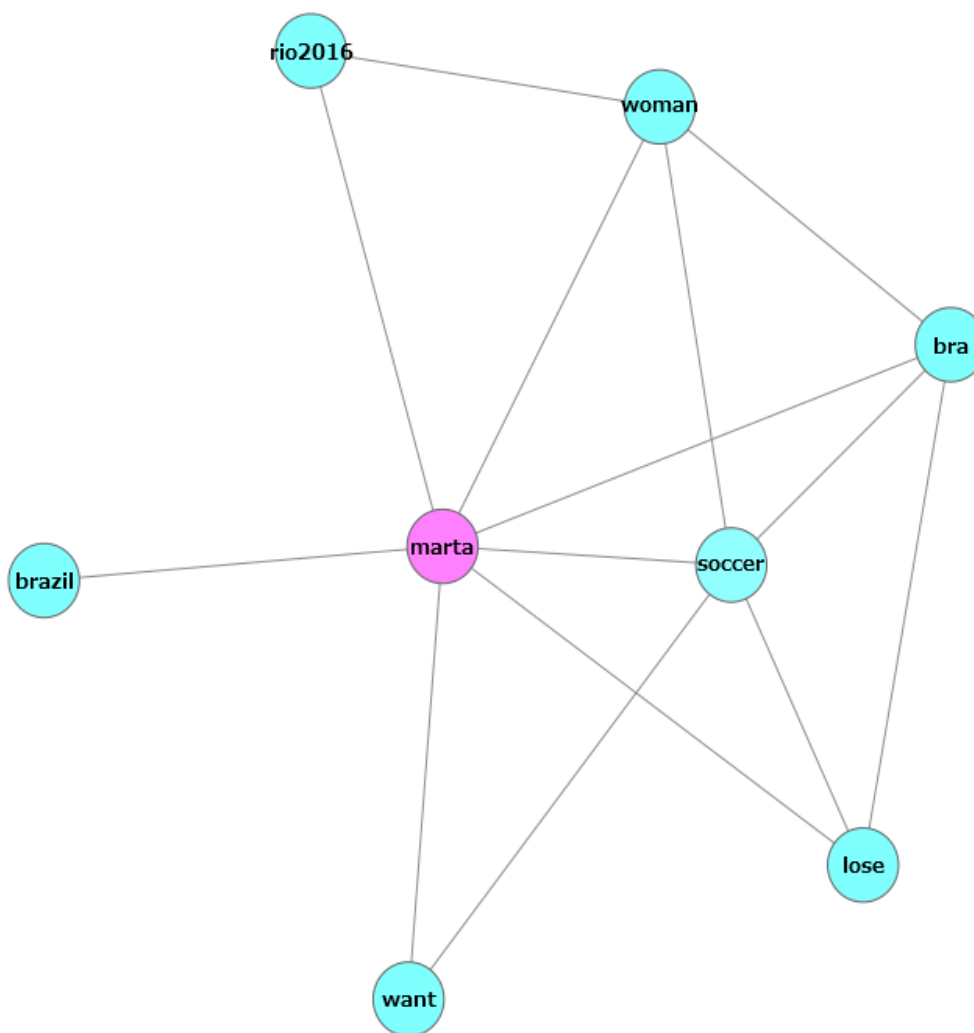


Figura 17 – Diagrama de coocorrência para atleta Marta Vieira da Silva de 16/08/2016 até 21/08/2016. Fonte Autor

4.6 Visão dos Desejo por idioma

Considerando todos os *tweets* marcados como desejo da língua inglesa, a maioria dos desejos possui sentimento neutro, seguido por positivo e minoria negativo como mostra a Figura 18 e Tabela 13.

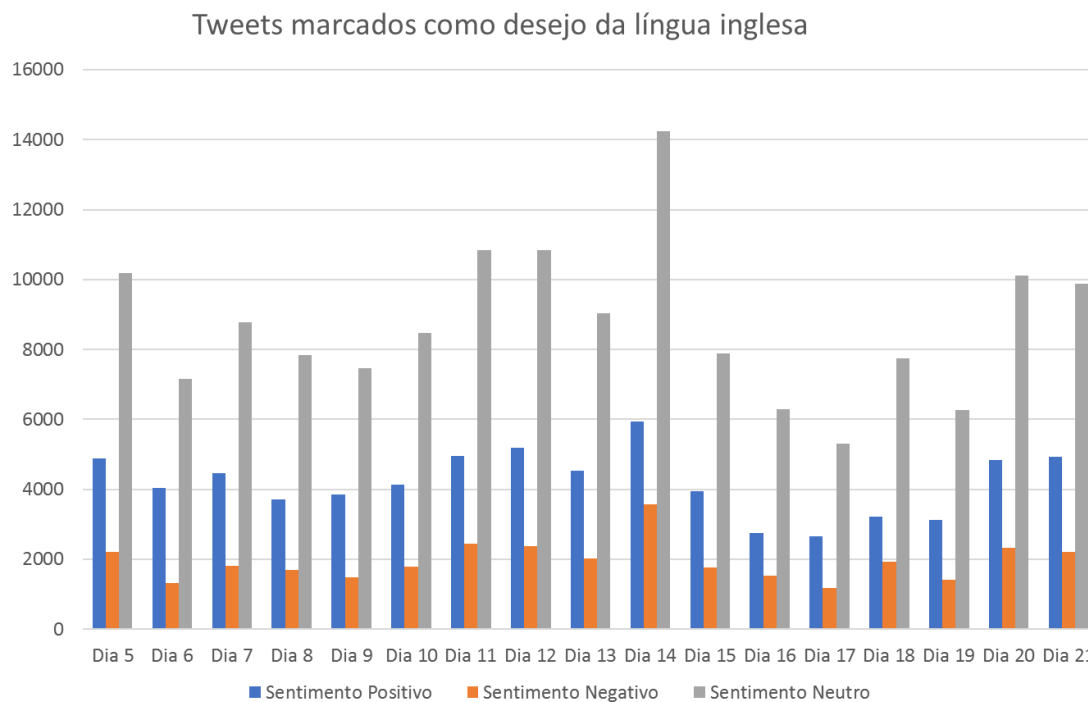


Figura 18 – Gráfico quantitativo da classificação positivo, negativo e neutro de tweets marcados como desejo para tweets da língua inglesa. Fonte: Autor

Tabela 13 – Tabela quantitativo da classificação positivo, negativo e neutro de tweets marcados como desejo para tweets da língua inglesa. Fonte: Autor

	Sentimento Positivo	Sentimento Negativo	Sentimento Neutro
Dia 5	4892	2205	10181
Dia 6	4028	1326	7158
Dia 7	4464	1807	8783
Dia 8	3709	1685	7845
Dia 9	3856	1492	7453
Dia 10	4141	1786	8464
Dia 11	4959	2435	10842
Dia 12	5179	2365	10840
Dia 13	4521	2026	9043
Dia 14	5937	3562	14249
Dia 15	3937	1765	7875
Dia 16	2752	1534	6299
Dia 17	2651	1188	5305
Dia 18	3224	1934	7740
Dia 19	3132	1404	6267
Dia 20	4844	2335	10123
Dia 21	4933	2211	9867

Gerando o gráfico normalizado com apenas o sentimento positivo e negativo é possível perceber que na língua inglesa os desejos com sentimento positivo apresentaram maior constância que os negativos.

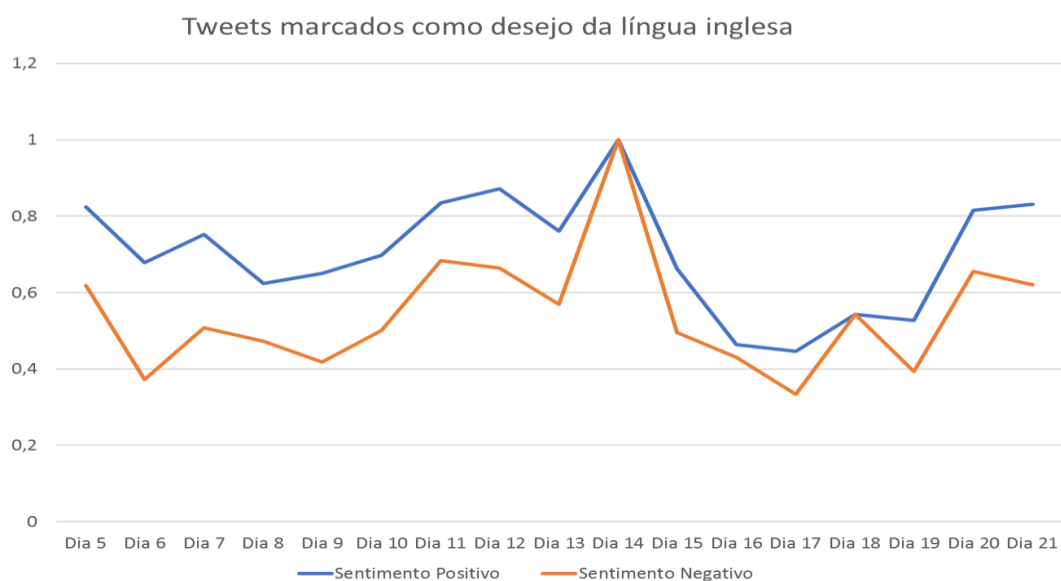


Figura 19 – Gráfico normalizado da classificação positivo e negativo de tweets marcados como desejo para tweets da língua inglesa. O valor 1 no eixo Y significa 100% Fonte: Autor

Considerando todos os *tweets* marcados como desejo da língua portuguesa, a maioria dos desejos possui sentimento neutro, seguido por positivo e minoria negativo como mostra a Figura 20 e Tabela 14.

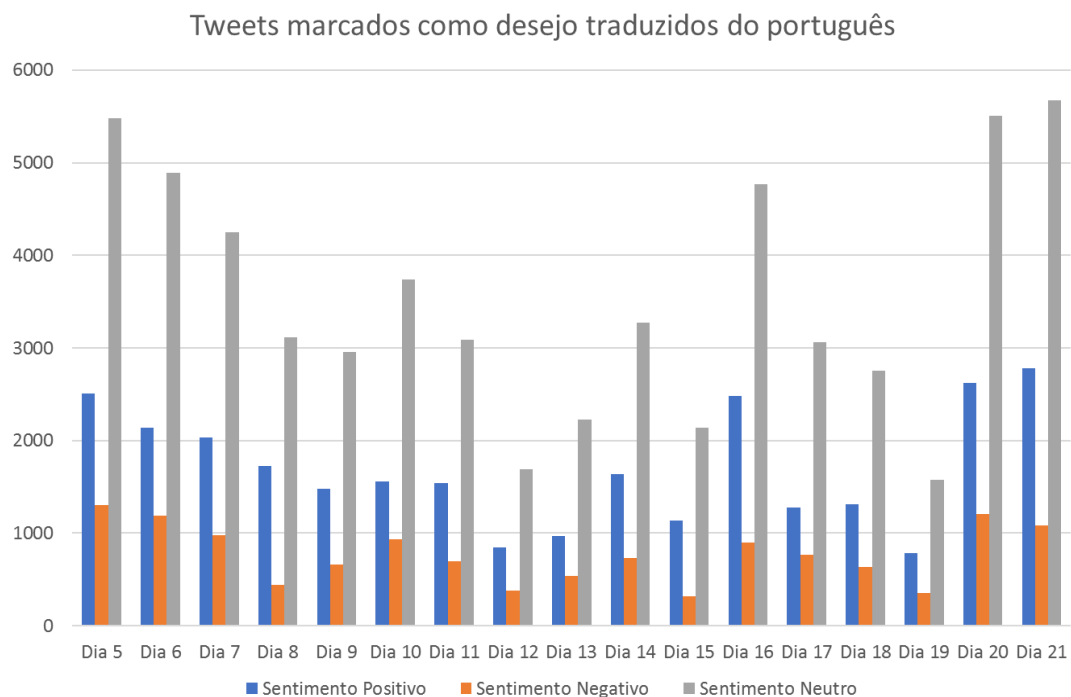


Figura 20 – Gráfico quantitativo da classificação positivo, negativo e neutro de tweets marcados como desejo para tweets da língua portuguesa. Fonte: Autor

Tabela 14 – Tabela quantitativo da classificação positivo, negativo e neutro de tweets marcados como desejo para tweets da língua portuguesa. Fonte: Autor

	Sentimento Positivo	Sentimento Negativo	Sentimento Neutro
Dia 5	2506	1299	5477
Dia 6	2136	1191	4892
Dia 7	2034	980	4252
Dia 8	1726	439	3117
Dia 9	1475	661	2953
Dia 10	1557	934	3739
Dia 11	1544	692	3090
Dia 12	846	379	1694
Dia 13	971	542	2225
Dia 14	1634	732	3271
Dia 15	1136	322	2143
Dia 16	2482	900	4770
Dia 17	1277	766	3066
Dia 18	1316	634	2750
Dia 19	786	352	1573
Dia 20	2619	1207	5506
Dia 21	2778	1082	5673

Gerando o gráfico normalizado com apenas o sentimento positivo e negativo é possível perceber que na língua portuguesa os desejos positivos e negativos foram mais inconstantes em relação à língua inglesa. Vale ressaltar que o máximo negativo acontece no primeiro dia dos Jogos e máximo positivo no último dia dos Jogos.

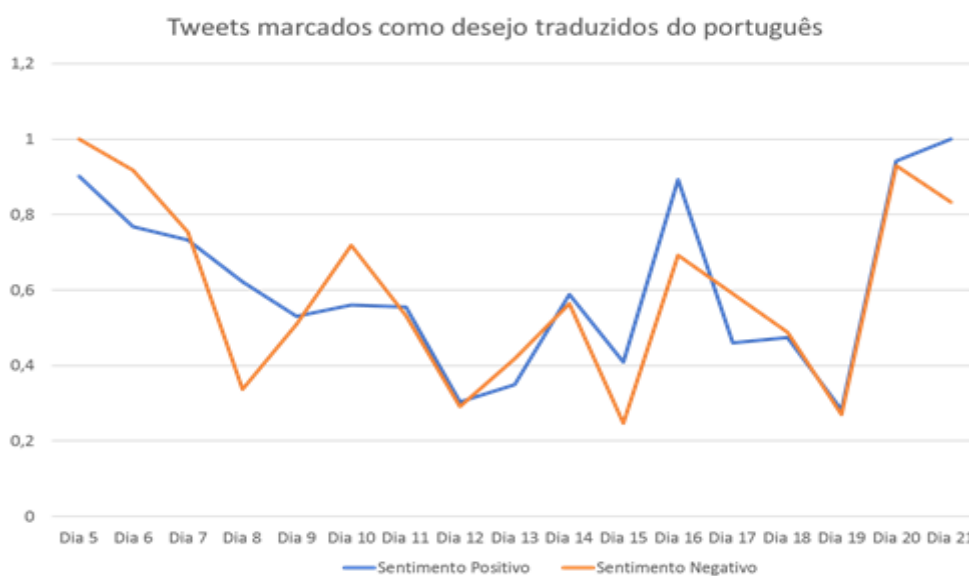


Figura 21 – Gráfico normalizado da classificação positivo e negativo de tweets marcados como desejo para tweets da língua portuguesa. O valor 1 no eixo Y significa 100% Fonte: Autor

É possível perceber que a maioria dos *tweets* marcados como desejo em português e inglês foram classificados como neutros seguido de positivo e minoria negativo.

4.7 Visão de Resultados Interessantes

No objetivo inicial foi pressuposto que na #Rio2016 seriam encontrados apenas desejos para atletas por se tratar de uma *hashtag* oficial dos Jogos Olímpicos. Mas em certos momentos dos Jogos alguns influenciadores foram mais citados que os atletas.

No dia 05/08/2016, cerimônia de abertura das olimpíadas, a artista Anitta foi citada em 415 *tweets* classificados como desejo, sendo 129 marcados como positivo, 70 marcados como negativo e 216 marcados como neutro. A Figura 22 mostra o diagrama formado a partir desses 415 *tweets*. A artista participou da cerimônia de abertura das olimpíadas e teve grande visualização. Enquanto na língua portuguesa ela foi citada em 415 *tweets* marcados como desejo na língua inglesa foi citada em apenas 29 *tweets*. A artista é brasileira o que leva a possibilidade que a nacionalidade junto com a língua que é falada no país da nacionalidade tem relação com desejos feitos em determinada língua.

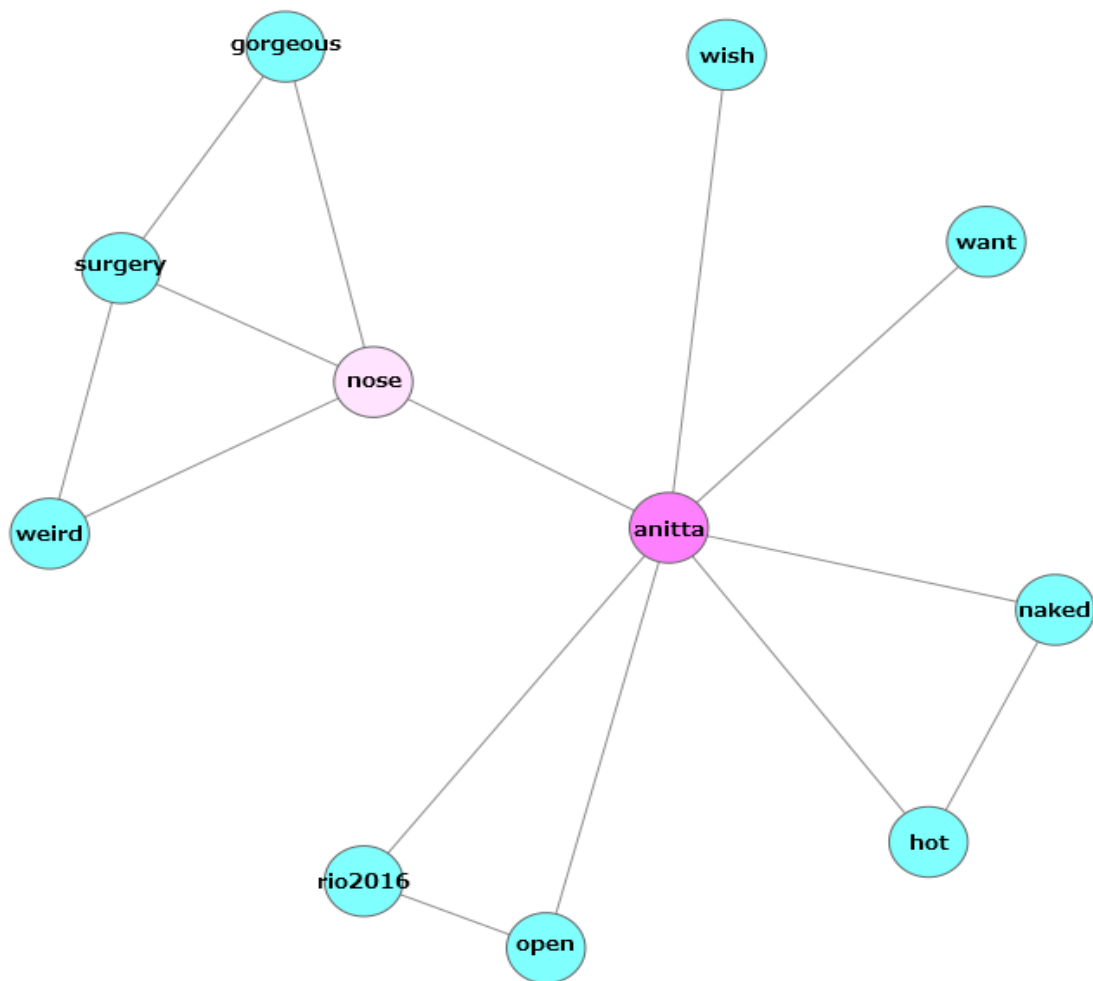


Figura 22 – Diagrama de coocorrência para artista Anitta de 05/08/2016. Fonte: Autor

No dia 05/08/2016, cerimônia de abertura das olimpíadas, o candidato à presidência americana Donald Trump foi citado em 632 *tweets* marcados como desejo na língua inglesa. A Figura 23 é o diagrama gerado a partir desses 632 *tweets*.

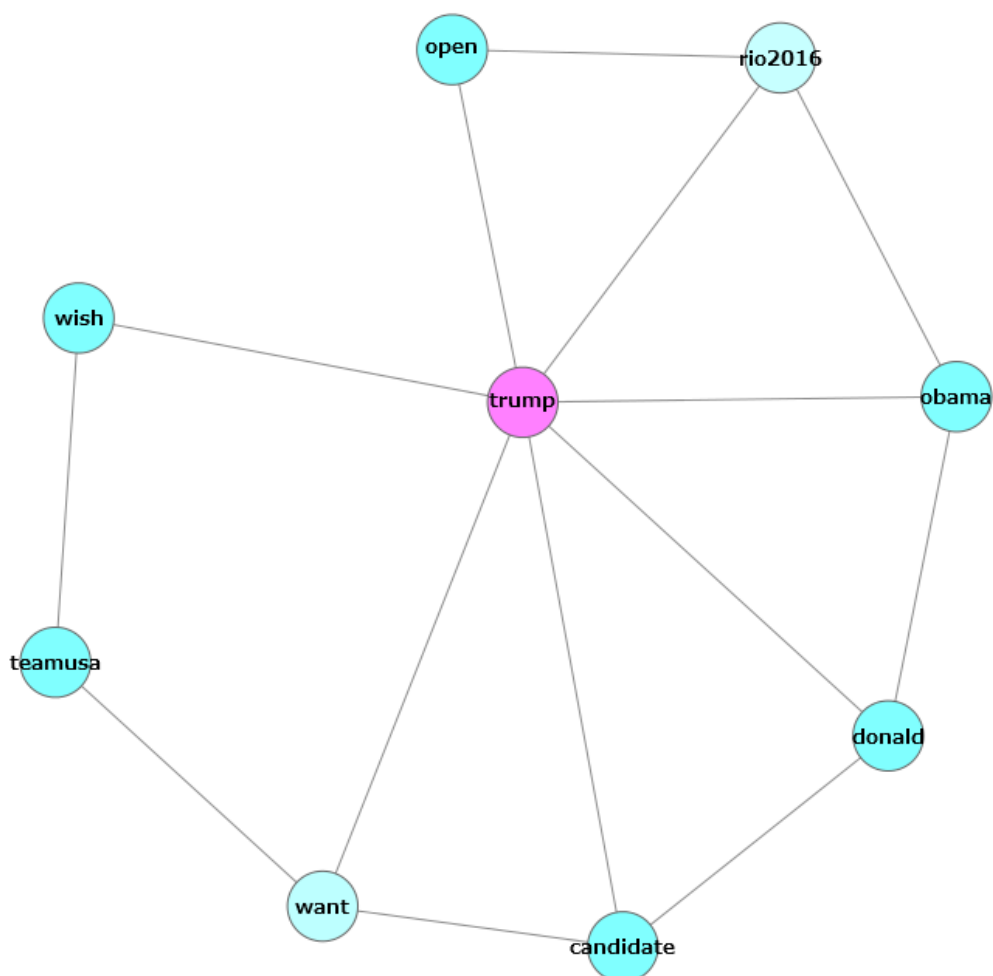


Figura 23 – Diagrama de coocorrência para candidato Donald Trump de 05/08/2016. Fonte: Autor

No mesmo dia 05/08/2016 o presidente da república Michel Temer recebeu 469 *tweets* marcados como desejo na língua portuguesa. A Figura 24 é o diagrama gerado a partir desses 469 *tweets*. Em inglês e em português dois políticos, o candidato à

presidência e o presidente da república foram citados em muitos *tweets* marcados como desejo. Isso mostra a possível relação entre política e o desejo das pessoas em relação a seus líderes.

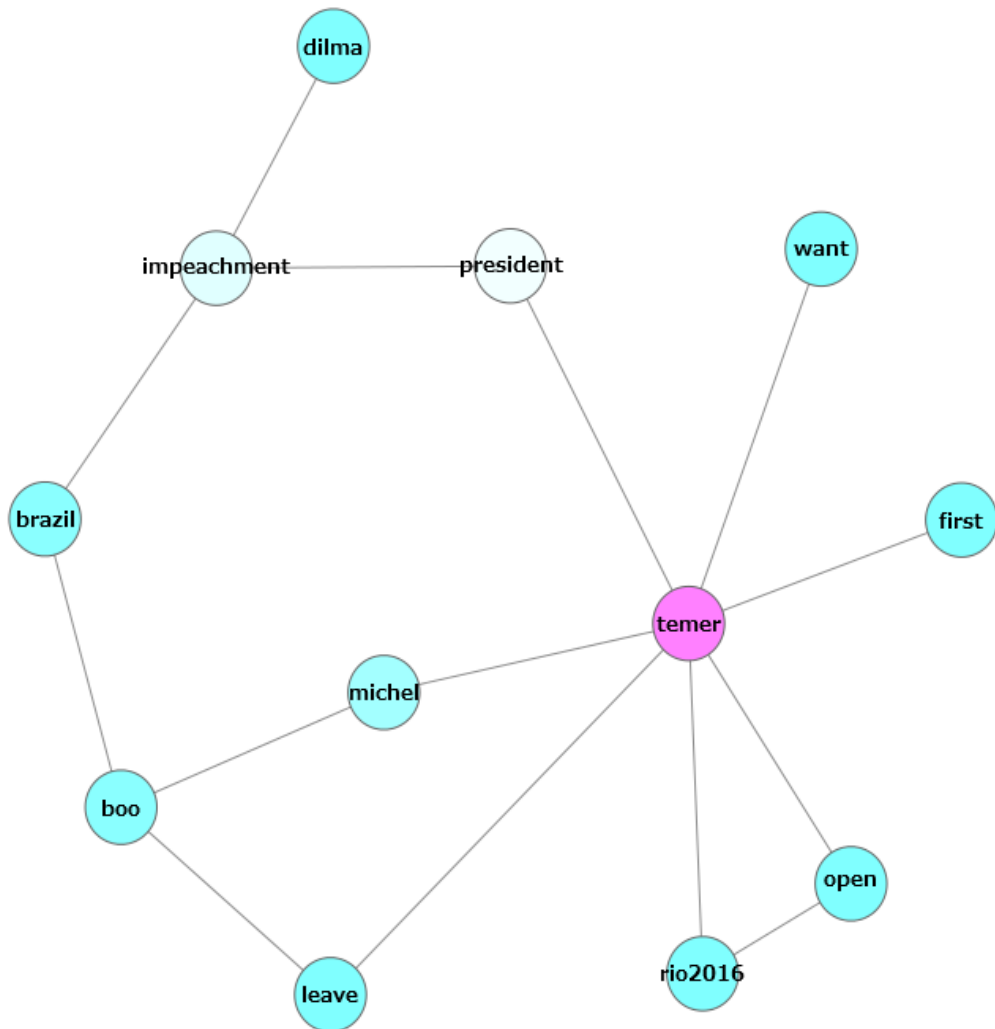


Figura 24 – Diagrama de coocorrência para presidente Michel Temer de 05/08/2016. Fonte: Autor

No dia 21/08/2016, dia da cerimônia de encerramento das olimpíadas, o narrador Galvão Bueno foi citado em 304 *tweets* marcados como desejo. A Figura 25 é o diagrama gerado a partir desses 304 *tweets*. Nesse diagrama aparece a emissora de TV Rede Globo. Galvão Bueno narrava nesse dia a cerimônia de encerramento dos Jogos Olímpicos e a final masculina do futebol. Meios de comunicação também mostram possível relação com o sentimento de desejo.

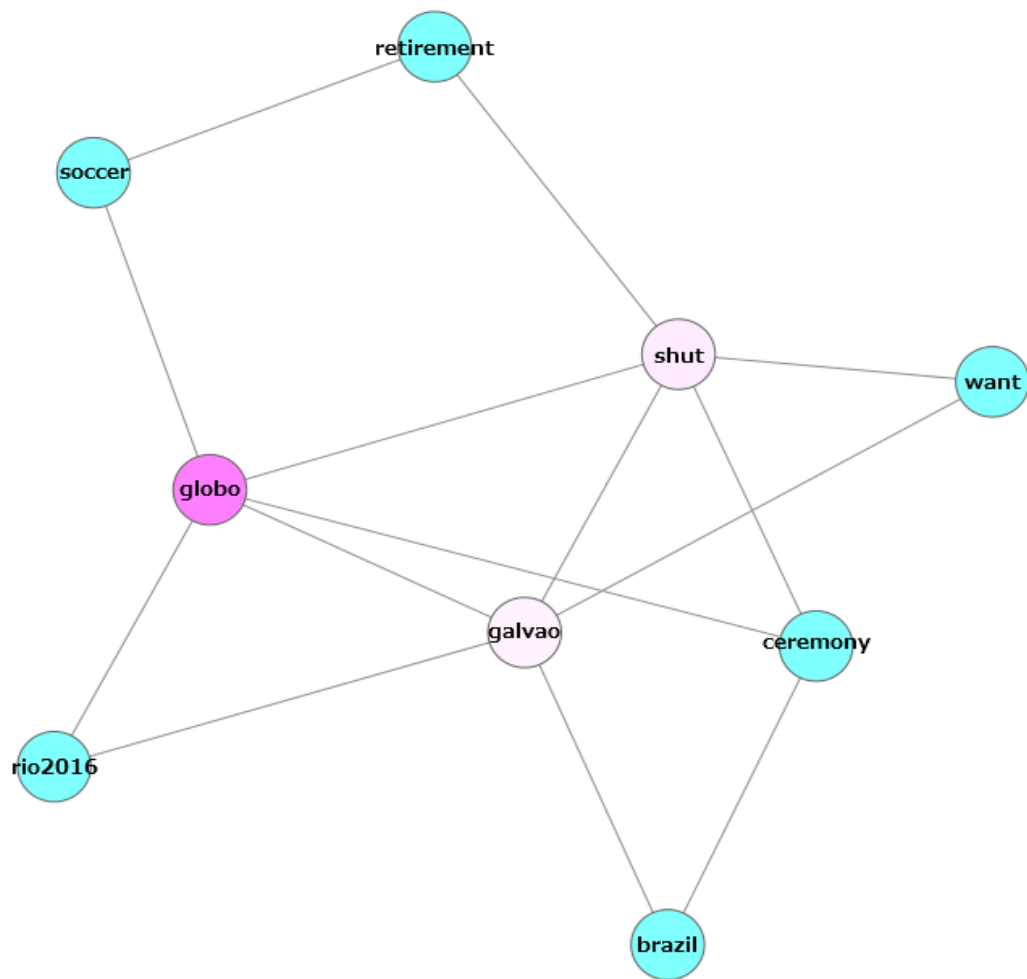


Figura 25 – Diagrama de coocorrência para narrador Galvão Bueno de 21/08/2016. Fonte: Autor

5 Conclusão

Os Jogos Olímpicos de 2016 produziu grande quantidade de dados sobre desejos em Português e Inglês. A combinação das técnicas de processamento de linguagem natural (NLP), mineração de textos, análise de sentimentos, e redes de coocorrência foi capaz de detectar desejos e levantar possíveis motivos para os mesmos. Foi possível classificar e quantificar os desejos em positivo, negativo e neutro o que ajuda no entendimento de como desejos funcionam. O melhor entendimento do funcionamento de desejos pode ajudar criação de métodos que possam prever situações de causa e consequências que levam a um desejo.

O processo de descoberta de desejo se mostrou capaz de mapear os principais motivos que levam a um desejo positivo ou negativo para atletas e para influenciadores. O processo utilizou as redes de coocorrência para conseguir encontrar a coocorrência das principais palavras relacionadas a desejos positivos e negativos. O processo se mostrou eficiente, mas com resultados sensíveis a três variáveis: Objeto alvo estudado, período do estudo e sentimento positivo, neutro ou negativo estudado. Fatores de desejo não relacionado a atletas também foram observados como atenção da mídia, política e nacionalidade. Principais verbos relacionados aos desejos realizados durante os Jogos Olímpicos também conseguiram ser levantados. Foi observado que os dias em que os atletas competiram foram os dias em que os fãs mais se expressaram nas redes sociais, isso mostra uma relação entre o sentimento de desejo e o sentimento de torcida.

Foi observado que desejos polarizados surgem de fatores como expectativas criadas

por fãs para atletas. A expectativa para cada atleta é individual e dependente de seus comportamentos, características e história com seus fãs. Em uma tendência geral, desejos positivos para atletas podem ser observados quando um atleta está de forma contínua atendendo as expectativas feitas sobre ele. Desejos negativos podem ser observados quando um atleta não atende apenas uma expectativa feita sobre ele. Quando um atleta faz algo que não agrada seus fãs a relação entre desejos positivos e negativos pode mudar e se manter alterada por pelo menos alguns dias.

O conjunto de atletas mostrou-se ótimo objeto alvo material para estudo de desejos por conseguir gerar resultados tão particulares para cada atleta. O Twitter mostrou-se uma ótima rede social para pesquisa por ser capaz de endereçar assuntos tão bem com o uso de *hashtags*. O Umigo mostrou-se mais focado na corretude da classificação quanto aos sentimentos polarizados do que quanto a cobertura da classificação, de forma que caso haja uma complexa junção de diferentes sentimentos na mesma frase o algoritmo tende a classificar como neutro do que assumir que a frase é polarizada. Dessa forma boa parte dos desejos é classificado como neutro, mas considerando apenas os pólos positivo e negativo é possível perceber que existem mais desejos positivos do que desejos negativos.

De modo geral, o trabalho conseguiu gerar conhecimento para enriquecer o debate sobre o sentimento de desejo com tanta qualidade quanto trabalhos das áreas de Psicologia e Filosofia.

6 Referências Bibliográficas

- [1] DESTRÉE, Pierre. ACRASIA ENTRE ARISTÓTELES E SÓCRATES, Revista Analytica, Rio de Janeiro, v.8, n.2, p135-164, 2004.
- [2] C.BIRCKHOUSE, Thomas, D.SMITH, Nicholas. “Os paradoxos socráticos”, Livro Platão, H. BENSON, Hugh, Brasil, Artmed Editora, p248-260, 2006.
- [3] GONÇALVES, J. C., SANTORO, F., BAIÃO, F., REVOREDO, K. “Discovering Intentions and Desires within Knowledge Intensive Processes”, Business Process Management Workshops, Springer Internatinal Publishing Switzerland, Austria, p273-284, 2015.
- [4] HAGEN, C. R., RATZ, D., POVALEJ, R. “Towards self-organizing knowledge intensive processes”, Journal of universal knowledge management, v.0, n. 2, p.148-169, 2005.
- [5] MANNING, Christopher D, SURDEANU, Mihai, BAUER, John, FINKEL, Jenny, BETHARD, Steven J., McClosky, David. “The Stanford CoreNLP Natural Language Processing Toolkit”, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, p55-60, 2014
- [6] RIBEIRO, Filipe N., ARAÚJO, Matheus, GONÇALVES, Pollyanna, GONÇALVES, Marcos André, BENEVENUTO, Fabrício. “SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods”, EPJ Data Science, SpringerOpen Journal, 2016.

[7] Twitter, documento disponível em:

https://blog.twitter.com/official/pt_br/a/pt/2016/rio2016-a-emo-o-dos-jogos-ol-mpicos-no-twitter.html

[8] FELDMAN, Ronen, SENGER, James. Livro The Text Mining Handbook, New York, Cambridge University Press, 2006, ISBN 978-0-521-83657-9.

[9] MANNING, Christopher D., RAGHAVAN, Prabhakar, SCHÜTZE, Hinrich, Livro Introduction to Information Retrieval, New York, Cambridge University Press, 2008, ISBN 052186571-9.

[10] FRANÇA, J., NETTO, J., CARVALHO, J., SANTORO, F., BAIÃO, F., PIMENTEL, M. “KIPO: The Knowledge Intensive Process Ontology”, Software & Systems Modeling April 2014, Springer International Publishing Switzerland, 2014.

[11] RISH, Irina. “An empirical study of the naive Bayes classifier”, IJCAI 2001 workshop on empirical methods in artificial intelligence, IBM Press, New York, p41-46, 2001.

[12] SANTORINI, Beatrice. "Part-of-speech tagging guidelines for the Penn Treebank Project.", Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, 1990.

[13] TOUTANOVA, Kristina, MANNING, Christopher D. “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger.”, Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very

Large Corpora (EMNLP/VLC-2000), p63-70, 2000.

[14] FINKEL, Jenny Rose, GRENAGER, Trond, MANNING, Christopher. “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”, Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), p363-370, 2005.

[15] PANG, B., LEE, L., “A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts”, Proceedings of ACL-04, 2004.

[16] LI, Shoushan, LEE, Sophia Yat Mei, CHEN, Ying, HUANG, Chu-Ren, ZHOU, Guodong. “Sentiment Classification and Polarity Shifting”, Proceeding COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, p635-643, 2010.

[17] LEVALLOIS, Clement. “Umigon: sentiment analysis on Tweets based on terms lists and heuristics”, Proceedings 7th International Workshop on Semantic Evaluation (SemEval 2013), p414-417, 2013.

[18] HIGUCHI, Koichi, Livro KH Coder 3 Reference Manual, 2016.

[19] FRUCHTERMAN, T. M. J., REINGOLD, E. M., “Graph Drawing by Force-directed Placement,” Software - Practice and Experience, 1991.

[20] Twitter, documento disponível em: <https://dev.twitter.com/overview/api>

[21] WU, Yonghui, SCHUSTER, Mike, CHEN Zhifeng, LE, Quoc V., NOROUZI

Mohammad, “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”, Google Research Blog, 2016.

[22] AutoIt documento disponível em: <https://www.autoitscript.com/site/autoit/>

[23] GOLDBERG, Andrew B., FILLMORE, Nathanael, XU, David Andrzejewski Zhiting, GIBSON, Bryan, ZHU Xiaojin, “May All Your Wishes Come True: A Study of Wishes and How to Recognize Them”, Computer Sciences Department, University of Wisconsin-Madison, 2009.