INSTITUTO INFNET

NICOLAS DE SOUSA TEODOSIO E VICTOR HUGO NOVAIS RODRIGUES

ANÁLISE DE SENTIMENTO E MINERAÇÃO DE OPINIÕES APLICADO NO TWITTER

RIO DE JANEIRO

2016

INSTITUTO INFNET

NICOLAS DE SOUSA TEODOSIO E VICTOR HUGO NOVAIS RODRIGUES

ANÁLISE DE SENTIMENTO E MINERAÇÃO DE OPINIÕES APLICADO NO TWITTER

Trabalho de Conclusão de Curso apresentado ao Programa de Graduação em Engenharia da Computação do Instituto Infnet como parte dos requisitos necessários à obtenção do título de Bacharel em Engenharia da Computação.

Orientador:
CASSIUS FIGUEIREDO

RIO DE JANEIRO

2016

NICOLAS DE SOUSA TEODOSIO E VICTOR HUGO NOVAIS RODRIGUES

ANÁLISE DE SENTIMENTO E MINERAÇÃO DE DADOS APLICADO NO TWITTER

Trabalho de Conclusão de Curso apresentado ao Programa de Graduação em Engenharia da Computação do Instituto Infnet como parte dos requisitos necessários à obtenção do título de Bacharel em Engenharia da Computação

Aprovada em XX agosto de 2016.

BANCA EXAMINADORA

Profº. Cassius Figueired, M.Sc. - Orientador
Instituto INFNET

Profª. XXXX, titulacao.
Universidade

Profº. xxx, TITULACAO
Universidade

Rio de Janeiro
2016

À minha família.

Agradecimentos

Agradeço, inicialmente,

Resumo

Atualmente a internet e micro blogs em geral têm se tornado uma ferramenta de comunicação poderosa entre usuários de Internet. Bilhões de pessoas compartilham informações e opiniões todos os dias, fazendo desse espaço um ótimo campo de pesquisas comercias, acadêmicas e sociológicas. Como o fenômeno é relativamente recente – o Twitter foi criado apenas em 2006 – ainda existem poucas pesquisas destinadas ao tema.

Os principais desafios para aplicação dessa técnica estão relacionados a linguagens naturais sensíveis ao contexto que não trazem resultados satisfatórios quando utilizam-se modelos matemáticos muito simples, sendo necessário um grande investimento de tempo em aperfeiçoar os modelos matemáticos disponíveis e adaptá-los à solução em questão.

Outro desafio interessante é a aplicação de técnicas de mineração de opiniões no português, onde não existem muitos trabalhos relacionados e massas de treino disponíveis para consulta.

O objetivo deste trabalho é explorar o potencial existente em pesquisas de opinião que podem ser feitas através de análises nas comunicações feitas em língua portuguesa nas redes sociais todos os dias.

Palavras-chave: Análise de sentimento, mídias sociais, twitter, mineração de opiniões, processamento de linguagem natural, linguagens sensíveis a contexto, naive bayes.

Abstract

Palavras-chave: xxxxxxx.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

API Application Program Interface 1

PNL Processamento de Linguagem Natural 12

IA Inteligência Artificial 12

JSON *Javascript Object Notation* 19

Sumário

Capítulo 1

Introdução

Através do fenômeno da popularização da Internet vivemos hoje um período conhecido como "Era da conhecimento"[1]. Nesse contexto, redes sociais conhecidas, como Facebook e Twitter se tornaram bastante populares por permitirem a seus usuários acesso à um ambiente onde todos possuem voz e vez para se expressar e por consequência, para se informar sobre tudo que acontece no mundo. Através de Application Program Interface (API) disponibilizadas por essas redes sociais, possuímos fácil acesso a um grande volume de opiniões catalogadas - através de *hashtags* - que podem ser utilizadas em pesquisas de opinião sobre um tema ou assunto específico. Tal cenário apresenta-se como uma grande oportunidade de pesquisa em áreas acadêmicas, sociais e comerciais. Porém, quando o objeto de estudo é a língua portuguesa, nota-se que a mesma carece de trabalhos e implementações na área de mineração de opiniões e análise de sentimento (REFERÊNCIA). Alguns motivos explicam essa carência: poucos investimentos na área de ciência e engenharia da computação em nosso país e a grande dificuldade que a língua portuguesa apresenta ao ser interpretada através de processamento de linguagem natural. [2]

1.1 Motivação e Objetivos

1.2 Principais contribuições

1.3 Recursos utilizados

1.4 Organização do trabalho

Este trabalho está estruturado em 5 capítulos da seguinte forma: no Capítulo 2, para embasamento teórico, são apresentados os conceitos de (CONTINUA). Em seguida, no Capítulo 4 , é feita uma análise sobre os principais trabalhos relacionados ao uso dos ... . No Capítulo 2, os conceitos do arcabouço utilizado ... , são descritos. Nesse capítulo são mostrados os motivos para a escolha desse arcabouço, ... A proposta XXX é apresentada no Capitulo 4, onde a arquitetura da proposta é detalhada, assim como seus componentes e algoritmos. Em seguida, o Capítulo 5 apresenta as ferramentas utilizadas para implementação da proposta, o ambiente implementação, a descrição dos experimentos e os principais resultados obtidos com o XXX, assim como a análise dos valores encontrados. Por fim, o Capítulo 6 conclui este trabalho, ressaltando os objetivos alcançados com as propostas. As principais vantagens e desvantagens da proposta são discutidas, assim como alguns trabalhos futuros que podem ser desenvolvidos.

Capítulo 2

Referencial Teórico

2.1 Mineração de opinião

É de conhecimento comum que há um acúmulo de dados por toda a internet. Artigos, informações de usuários, comportamento de usuários, essas são alguns tipos de informação que podem ser encontrados hoje na internet. Esse grande acumulo não garante informações confiáveis ou uma análise correta sobre os dados, por isso hoje há uma grande urgência para novas teorias computacionais e ferramentas que ajudem a analisar essa quantidade de dados que só aumenta [3]. E dentro dessa enorme gama de dados, existem as informações adicionadas por usuários através de texto que remetem a suas reações a determinadas situações ou objetos.

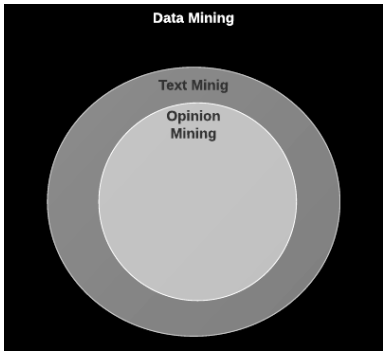

Figura 2.1: Diagrama de Venn - Mineração de Dados

2.1.1 Sentimento

De acordo com psicólogo Klaus R. Scherer, sentimento é um breve episódio da resposta sincronizada de todos os ou grande parte dos subsistemas orgânicos em resposta a um

evento interno ou externo de grande significância[4]. Algumas outras definições utilizadas são:

- Ato ou efeito de sentir;
- Aptidão para receber as impressões;
- Sensação, sensibilidade;
- Consciência íntima;
- Faculdade de compreender, intuição e percepção;

A mineração de opinião, também conhecida como mineração de sentimento, análise de sentimento ou extração de opinião, é um campo dentro da mineração de dados [5] que tem como objetivo extrair o sentimento do texto escrito por uma pessoa, sem a interferência humana durante o processo. Porém, existe dificuldade em afirmar categoricamente o que é sentimento.

2.1.2 Desafios

No campo de mineração de opinião, existem uma série de desafios que devem ser tidos como grandes pontos de atenção para quem deseja aplicar essa técnica de forma correta.

- Em blogs e redes sociais é comum encontrar textos com erros de ortografia ou escritos de forma informal, contendo gírias e abreviações comuns dentro da comunicação virtual;
- Dificuldade em discernir uma opinião ou um fato, especialmente quando existem opiniões embutidas em fatos;
- Os textos podem conter ironias e sarcarmos, que são especialmente difíceis de serem identificados e podem impactar os resultados;
- Um texto pode se referir à dois temas diferentes - política e ideologia, por exemplo - com opiniões diferentes sobre os mesmos, o que pode confundir a classificação;

2.1.3 Etapas

O processo de mineração de opinião consiste em 3 etapas: [6]

- Coleta de dados;
- Classificação;
- Análise dos resultados;

2.1.3.1 Coleta de dados

Nesta etapa é conduzida uma busca por opiniões nas mais diversas fontes que podem ser úteis: artigos, sites, comentários, anúncios dentre outras. Como explicado anteriormente, deve-se visar identificar se a informação coletada é uma opinião ou fato. Fatos podem ser descartados imediatamente, porém opiniões apresentadas através de fatos, podem ser úteis.

Existem diversas maneiras de coletar sistematicamente fontes para extrair e armazenar os dados que serão utilizados, dentre elas as mais famosas estão o desenvolvimento *crawlers* e a utilização de APIs.

2.1.3.2 Classificação

A classificação é a alma do processo de mineração de opinião. Nesta etapa é determinada a polaridade do objeto de estudo, pretendendo determinar se o mesmo é positiva, negativa ou neutra.

Essa etapa é a principal responsável pela acurácia da análise. Por ser a etapa mais delicada do processo é onde ocorrem a maior parte dos erros. Existem diversas técnicas e ferramentas que ajudam a mitigar tais problemas que serão abordadas mais adiante, no Capítulo 4.

2.1.3.3 Análise dos resultados

A análise dos resultados envolve cruzar as informações de polaridade obtidas através texto com qualquer outra informação que exista sobre quem produziu aquela opinião. Desta forma, é possível, por exemplo, determinar qual gênero - masculino ou feminino - tem uma maior aceitação à um produto ou personalidade. As possibilidades para cruzar os

dados e obter *insights* será proporcional a quantidade de informações coletadas durante o processo.

2.1.4 Aplicações práticas

Um algoritmo capaz de extrair opiniões de um texto pode ser aplicado em diversos cenários:

2.1.4.1 Pesquisa de opinião sobre um produto

Mineração de opinião pode ser usada por uma companhia para determinar se um certo produto lançado ao mercado atingiu a aceitação prevista, como forma de entender a percepção do público e guiar estrategicamente ações de marketing e relações públicas. Ainda é possível prospectar o sentimento associado a um produto antes mesmo do seu lançamento, visando antecipar *insights* que podem ser valiosos durante o seu desenvolvimento.

2.1.4.2 Análise sobre pessoas públicas

Da mesma forma, é possível utilizar a mesma técnica e direcionar as análises para uma personalidade pública. Por exemplo, é possível determinar a aceitação ou rejeição de um político durante o mandato ou período de eleições, gerando dados que podem ser decisivos na definição de suas estratégias de campanha.

2.1.4.3 Bolsa de valores

Os números do mercado financeiro são uma consequência direta do sentimento que pessoas(investidores) possuem sobre uma empresa [7]. A opinião extraída de especialistas e sites de notícias podem ser usados como um dos fatores decisivos para compra e venda de ações.

2.1.5 Fontes de dados

É notório que estamos rodeados de dados dentro da Internet, porém dentro do campo de minerações de opiniões, existem algumas fontes que se destacam pela abrangência e diversidade dos dados.

2.1.5.1 Mecanismos de busca

É possível utilizar mecanismos de busca para obter opiniões sobre praticamente qualquer temática. Este método possui uma particularidade: mecanismos de busca como Google e Bing destacam certas páginas de acordo com motivos desconhecidos, o que pode influenciar os resultados obtidos. De forma geral, essa análise é apenas um reflexo do que está sendo buscado naquele momento.

Um exemplo da utilização de mecanismos de busca para mineração de opinião é o site whatdoesinternetthink.net[8], que utiliza como base os mecanismos de busca Google e Bing para determinar a opinião sobre um tema específico ou comparar dois temas entre si.

2.1.5.2 Redes sociais

O intenso compartilhamento de informações o opiniões que vemos hoje nas redes sociais serve como uma excelente fonte de dados para a mineração de opiniões por dois motivos: diversidade e abundância. Somando-se os usuários de Facebook e Twitter por exemplo, obtemos uma amostra considerável da população mundial à disposição para pesquisas.

Para este trabalho, o Twitter foi escolhido como base para a coleta de dados, por ser uma rede social focada em opiniões de usuários e pela grande facilidade que existe em consumir os seus dados através da API pública disponibilizada pelo mesmo.

2.2 Twitter

Contando com uma base ativa de usuários que ultrapassa 300 milhões[9], o Twitter é conhecido como um *microblog* fundado em março de 2006 por Jack Dorsey, Evan Williams e Biz Stone. Após 10 anos de mercado, a empresa acumula números impressionantes: 300 bilhões de mensagens já foram compartilhadas por seus usuários, que em média enviam 500 milhões de *tweets*[10] - nome pelo qual as mensagens compartilhadas no microblog ficaram conhecidas na Internet - por dia. Os usuários trocam mensagens de até 140 caracteres[11] em um ambiente de rede social, que tem como objetivo dar à todos o poder de criar e compartilhar ideias e informações instantaneamente [9].

Dentro do Twitter, O usuário pode fazer uso de marcadores conhecidos como *hash-tags*[12], para vincular sua mensagem à um tópico específico. Apesar de simples, as *hashtags* pode ser usadas das mais diversas maneiras:

- Agrupar comentários e pensamentos acerca de um tema
- Estabelecer uma conexão entre dois tópicos
- Aproximar o usuários de um conteúdo relevante com auxílio de uma busca

2.2.1 Primavera Árabe

Um dos exemplos mais recentes e impressionantes de como as redes sociais desempenharam o papel de aproximar ideologias semelhantes e encorajar debates sociais profundos foi a Primavera Árabe - onda de manifestações e protestos que tiveram início em dezembro de 2010, tendo como cenário o Norte da África e Oriente Médio. Os principais alvos foram os regimes ditatoriais e patriarcais que há muito tempo estavam no poder.[13]. Redes sociais foram amplamente utilizadas para marcar encontros, debates e manifestações, além de mostrar para o mundo o que acontecia em tempo real, através do Twitter e outras redes sociais, como o YouTube.

Figura 2.2: O celular e a internet foram as armas dos rebeldes na Primavera Árabe. Fonte: Desconhecida

2.2.2 Análises de redes sociais

Este novo cenário possibilitou que a análise de redes sociais ganhasse incrível relevância nos campos de pesquisa social e comportamental[14]. Ao invés de analisar comportamentos individuais, atitudes e crenças, a análise de redes sociais foca sua atenção em entidades sociais ou atores interagindo entre si e como essas interações constituem uma estrutura que pode ser estudada e analisada.

Outro ponto levantado recorrentemente quando o assunto é análise de redes sociais é o como ela pode ser útil para estudos de ordem micro ou macro. No nível *micro*, as análises destinam-se a examinar díades, tríades ou outros pequenos sub-grupos. No nível *macro*, o objeto de estudo são grandes redes de atores sociais. Todos os dados obtidos durante a coleta permitem segmentar os atores sociais de diversas formas - gênero, idade, religião, posição demográfica, entre outros - possibilitando análises *micro* - a nível de apenas um usuário - ou *macro* - quando analisamos um conjunto de usuários. Por exemplo, os dados extraídos a partir da API do Twitter, que será abordada no Capítulo 3, nos permite entender como um usuário específico reagiu a uma *hashtag*. Da mesma forma, podemos olhar um cenário mais amplo, como por exemplo, todos usuários de uma região do país. As possibilidades de análise crescem e se tornam mais ricas conforme obtemos mais informações sobre os atores no momento de suas interações sociais.

2.3 API

Por definição formal, uma API é um conjunto de rotinas estabelecidos por um software para a utilização de suas funcionalidades e acessos à seus dados por outro software que não pretende entender sobre a sua implementação, apenas seus serviços. Através dessa interface, capaz de fazer uma abstração dos dados e funcionalidades de um software, conectar-se a estes serviços se torna muito mais fácil, para ambos os lados.

Outro ponto que demonstra a importância das APIs durante o desenvolvimento de software é a interoperabilidade. Atualmente, temos o mesmo serviço sendo oferecido em diferentes plataformas, como por exemplo *web*, *desktop*, *mobile*, entre outras. Cada plataforma possui características e implementações diferentes, porém é possível que todas as plataformas utilizem as APIs como meio único de acesso a dados e serviços, promovendo uma padronização de protocolos e funcionalidades e serviços, além de alta reusabilidade de código.

O Twitter, nossa fonte de dados durante este trabalho, possui uma API pública que pode ser utilizada por qualquer usuário da rede social [15]. Atualmente, é a API mais utilizada no mundo, com mais de 15 bilhões de requisições por dia, 3 vezes mais acionada do que as APIs do Google, segundo colocado no ranking.

Para efetuar uma comunicação eficiente com quem acessa à API, é necessário implementar um protocolo de acesso aos dados. Entre eles, os mais utilizados são os protocolos REST e SOAP.

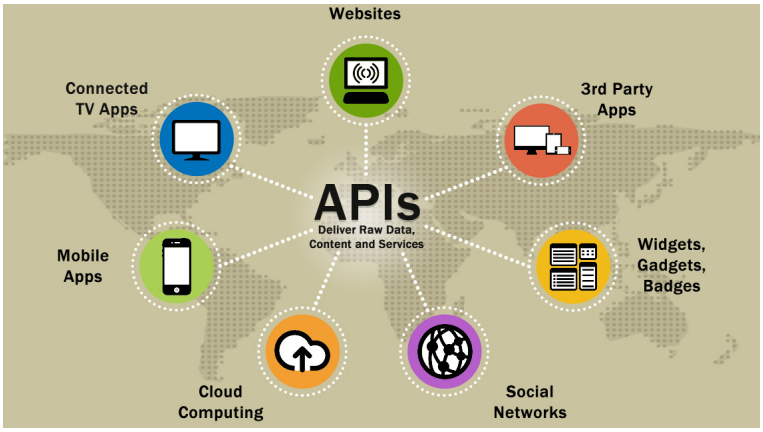

Figura 2.3: Papel das APIs integrando dados e serviços em diferentes plataformas. Fonte: http://www.programmableweb.com/

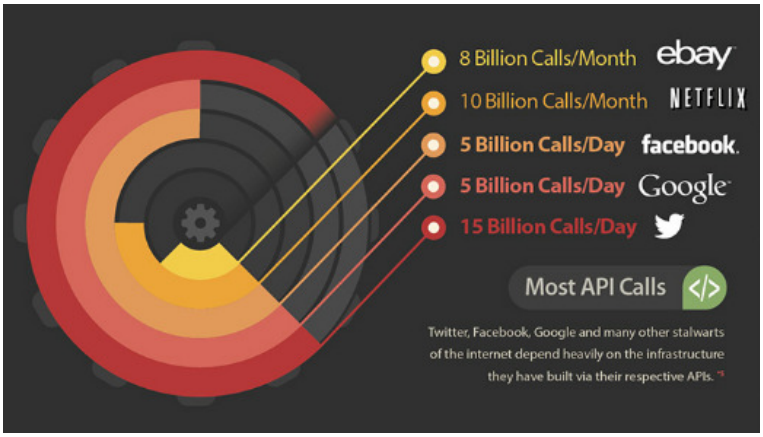

Figura 2.4: APIs mais utilizadas do mundo Fonte: SmartFile

2.3.1 REST

O protocolo REST foi criado em 2000 por Roy Fielding [16] durante sua dissertação de doutourado na *University of California Irvine*. Por ter sido criado dentro de um ambiente universitário, o objetivo do protocolo abraça a filosofia *open source*. Suas principais vantagens são:

- Segue a filosofia *open source*;
- Fácil implementação e manutenção;
- Separa claramente a implementação do cliente e do servidor;
- A comunicação não é controlada por uma entidade única;
- A informação pode ser armazenada pelo cliente previnindo múltiplas chamadas;

- Pode retornar a informação em múltiplos formatos (JSON, XML, entre outros);

Por outro lado, o protocolo REST possui algumas limitações. Entre elas, podemos destacar:

- Só funciona em cima do protocolo HTTP;
- Autorização e recursos de segurança devem ser implementados à parte;

Baseado nessas características, o protocolo REST é comumente utilizado para APIs de aplicações *Web* e *Mobile*, como por exemplo, a API do Twitter, LinkedIn e Slack.

2.3.2 SOAP

Criado em 1998 por Dave Winer et al com colaboração da Microsoft, o protocolo SOAP foca-se em endereçar necessidades do mercado corporativo. Como vantagem, o protocolo apresenta os seguintes aspectos:

- Segue uma abordagem mais formal, corporativa;
- Trabalha em cima de qualquer protocolo de comunicação, até mesmo assíncrono;
- Recursos de autorização e segurança incorporados de forma nativa;
- Pode ser descrito utilizando WSDL;

Entre suas principais desvantagens, podemos listar:

- Gasta-se muita banda trafegando metadados
- Difícil e implementação
- Pouco popular entre desenvolvedores *Web* e *Mobile*
- Retorna informação apenas em XML

Geralmente, o protocolo SOAP é mais utilizado em serviços financeiros, *gateways* de pagamento e serviços de telecomunicações.

2.4 Processamento de linguagem natural

2.4.1 Definição

Processamento de Linguagem Natural (PNL) baseia-se em modelos computacionais capazes de executar tarefas envolvem processar informações expresas em língua natural, como por exemplo, interpretação e tradução de textos. [17].

A pesquisa na área está voltada a três aspectos da comunicação essenciais:

- fonologia: estudo dos sons;
- morfologia: estudo da estrutura das palavras;
- semântica: estudo do significado;
- pragmática: estudo do significado aplicado a um contexto;

Neste trabalho, focaremos apenas no PNL aplicado à area da semântica e pragmática, responsável por estudar os elementos usados durante uma comunicação para se expressar através da língua (semântica) e a diversidade que pode surgir a partir de um contexto (pragmática). É também um estudo sobre como usuários de uma língua adquirem conhecimento sobre a mesma, através da comunicação oral ou escrita e como essa língua se altera ao longo do tempo.

Um dos grandes desafios da área é modelar o processamento de uma máquina para compreender uma estrutura tão complexa como uma linguagem. Existe um teste famoso na área de computação, o Teste de Turing, que levanta a questão "As máquinas podem pensar?". O artigo fundamenta conceitos chave sobre a Inteligência Artificial (IA), que serve como base para o PNL.

2.4.2 Teste de Turing

Introduzido pelo matemático britânico Alan Turing em seu artigo de 1950 "*Computing Machinery and Intelligence* [18], o Teste de Turing explora a capacidade de um computador demonstrar comportamento inteligente equivalente ou indistiguível dos seres humanos.

O teste é composto por três elementos: dois seres humanos, sendo um participante e um juiz e um computador.

O juiz conversa em linguagem natural com um outro ser humano e uma máquina através de um canal de texto, composto por um teclado e uma tela que renderiza a conversa. Todos os participantes estão em ambientes separados. O juiz deve ser capaz de distinguir a máquina do ser humano, caso contrário, a máquina é considerada bem sucedida no teste. O objetivo não é analisar se a máquina é capaz de responder corretamente e sim dizer quão próximas as respostas da máquina foram das do ser humano.

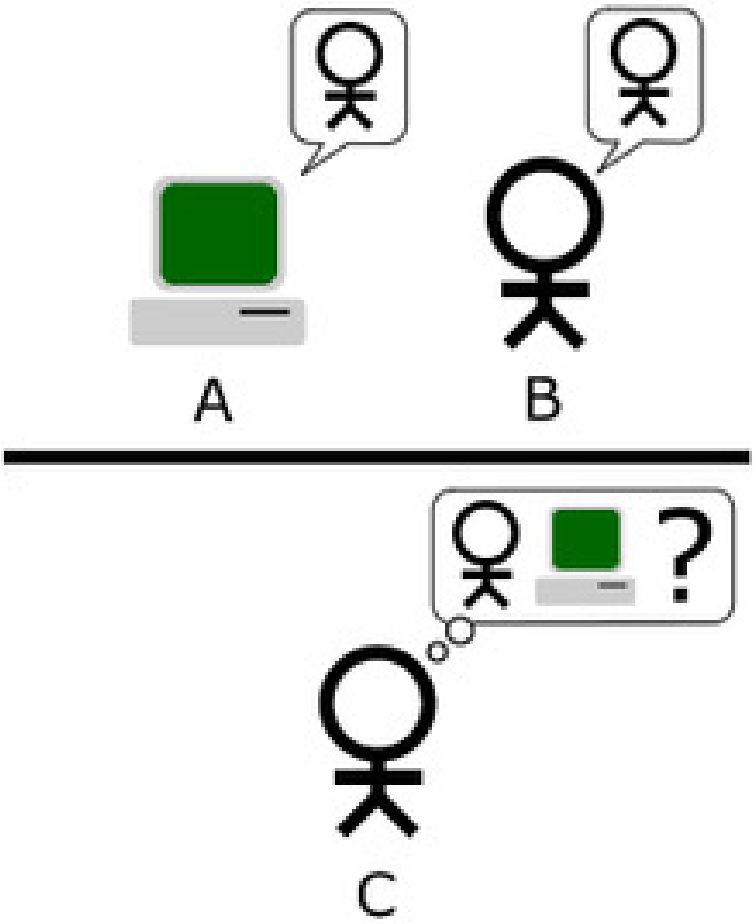

Figura 2.5: O participante A (máquina) e o participante B (humano) se comunicam por texto com o participante C (juiz). Fonte: Wikipédia

2.5 Classificador Naive Bayes

* Demonstração matemática do algoritmo * Uso dele em analise de sentimento/classificação

O classificador conhecido como *Naive Bayes* é algoritmos probabilístico baseado no Teorema de Bayes que não considera dependências que possam existir. Por este motivo, suas suposições são nomeadas "ingênuas", o que lhe confere uma maior simplicidade e um

desempenho maior, em relação a outros algoritmos de classificação [19]. É um método popular para categorização de textos, como por exemplo a classificação de *e-mails* em legítimos os *spam* [20]

2.5.1 O Teorema de Bayes

onde se infere qual é a probabilidade de um evento A dado um evento B e pode ser expressado pela seguinte equação:

$$P(A \mid B) = \frac{P(B \mid A)\ P(A)}{P(B)},$$

onde A e B são eventos e P(B) != 0.

- P(A) e P(B) são probabilidades de A e B sem considerar a relação entre ambos;
- P(A|B), uma probabilidade condicional, é a probabilidade de observar o evento A, dado que B ocorreu.
- P(B|A) é a probabilidade de observar o evento B, dado que A ocorreu.

Podemos também expressar a equação em puro português, como a seguir:

$$posterpriori = \frac{priori \times possibilidade}{evidência}$$

Suponha que queremos saber a probabilidade de um indivíduo possuir uma câncer, sem saber nada sobre o indivíduo. Porém, sabemos que a chance de um indivíduo estar infectado com tal câncer é de 1%, ou seja, uma probabilidade a priori. Em seguida, suponha que esta pessoa tenha 70 anos de idade e que essa probabilidade é de 0,2% e que 0,5% das pessoas doente possuem 70 anos de idade. Se assumirmos que a incidência de câncer e a idade estão relacionadas, podemos utilizar esta informação para melhor medir as chances desta pessoa estar doente. Logo, queremos saber a probabilidade de uma pessoa estar doente quando a mesma possui 70 anos de idade, ou probabilidade a posteriori.

$$(0,5\% \times 1\%) \div 0,2\% = 2,5\%$$

Portanto, o resultado do teorema demonstra que possuir 70 anos de idade aumenta a chance de uma pessoa possuir câncer, apesar desta probabilidade ainda ser baixa.

2.5.2 Aplicação neste trabalho

Nesse trabalho é utilizado o *Naive Bayes* para categorizar *tweets* extraídos do Twitter em positivos, neutros ou negativos. A diferença deste algoritmo para o Teorema de Bayes é assumir que a posição das palavras - eventos da probabilidade - que aparecem no texto não importa para determinar o resultado final.

Como visto em [21] o algoritmo computa qual a probabilidade de uma frase, denominada de documento pertencer a uma determinada classe(polaridade) *P(c/d)*, a partir da probabilidade a priori de *P(c)* do documento pertencer a esta classe e da probabilidades condicionais de cada termo *tk* ocorrer em um documento da mesma classe. O algoritmo tem como objetivo encontrar a melhor classe para um documento maximizando a probabilidade a posteriori conforme a equação abaixo, onde n_d é o número de termos no documento *d*.

$$C_{map} = argmax_{c \epsilon C} P(c|d) = argmax_{c \epsilon C} P(c) \prod 1sksn_d P(t_k/d)$$

Capítulo 3

Trabalhos relacionados

Com a crescente popularidade de blogs e redes sociais, os campos de mineração de opinião e análise de sentimento se tornaram objeto de estudo de alguns pesquisadores. Uma abordagem ampla sobre o assunto foi apresentada em Pang e Lee [22]. Em seu trabalho, os autores descrevem diversas técnicas e abordagens aplicáveis em sistemas orientados à informação. Entre as diversas aplicações sugeridas, destacam-se abordagens que visam substituir sites especializados em resenhas e recomendações, propondo que sistemas possam buscar opiniões de usuários de forma proativa ao invés de esperar que o mesmo exponha seu parecer através da solicitação do preenchimento de um formulário de pesquisa, resenha ou comentário. Tal abordagem pode ser aplicada para pesquisas de opinião sobre produtos, pessoas e serviços.

Em Gomes [23] a mineração de texto é aplicada em busca de notícias sobre economia de Portugal. O trabalho concentra-se em monitorar sites relevantes que abordam notícias sobre a economia do país para representar o sentimento expresso no texto, através dos títulos das reportagens.

Em Pak e Paroubek [24] o Twitter é utilizado como fonte dados para análises de sentimento. O idioma de estudo escolhido foi o inglês, mas grande parte das técnicas apresentadas podem ser aplicadas em outras línguas, visto que a coleta de dados e os algoritmos de classificação continuam inalteradas caso o objeto de estudo seja outro idioma.

Alguns trabalhos utilizam o português brasileiro como objeto de estudo, como por exemplo Tortella e Coelho [25]. Outros, se propõem a estudar um evento ou acontecimento finito, como por exemplo as eleições presidenciais no Brasil no ano de 2010 [26], os protestos populares contra a corrupção ocorridos em 2013 [27] e a Copa do Mundo da FIFA Brasil 2014 [28]. Nesses casos, *tweets* postados por usuários contendo *hashtags*

referentes ao evento a ser estudado são monitorados e salvos numa base de dados ao longo do evento. Após o fim do mesmo, os dados são classificados utilizando um algoritmo previamente treinado e os resultados são analisados afim de determinar a relevância, impacto e opiniões geral sobre de acordo com a opinião dos usuários.

Neste trabalho, será aplicado o processo de mineração de opinião e análise de sentimento de forma semelhante a Pak e Paroubek [24], porém utilizando português brasileiro como idioma de estudo. O objetivo é demonstrar como a mineração de opinião e análise de sentimento podem ser abordadas de forma abrangente, com a aplicação de técnicas generalistas. Além disso, é proposto um estudo de caso sobre a cerimônia do Oscar no ano de 2016 para também mostrar como um evento específico pode ser estudado, aplicando técnicas mais específicas que tem como objetivo tornar a classificação mais especializada e precisa, similar ao que foi feito em [26] [27] [28] .

Capítulo 4

Proposta

Como visto no Capítulo 2, existe uma corrente dentro da Mineração de Opinião que vem desenvolvendo maneiras de explorar o conteúdo digital gerado pela nossa sociedade todos os dias em redes sociais, através de técnicas utilizando Processamento de Linguagem Natural e *Machine Learning*, principalmente. Com este fato surge a oportunidade de explorar novas ferramentas na solução de problemas que envolvem pesquisas de opinião de forma geral. Neste trabalho propõem-se um *framework* que torna possível fazer pesquisas de opiniões em língua portuguesa sobre qualquer tema que seja rastreável a partir de uma *hashtag* no Twitter. Para tal é necessário que o framework criado seja capaz de:

1. Coletar *tweets* escritos em língua portuguesa que contenham uma determinada hashtag;
2. Armazenar as mensagens em uma base de dados;
3. Classificar as mensagens de acordo com a polaridade: negativo, neutro e positivo;
4. Extrair *insights* que auxiliem a tomada de decisão a partir da massa de dados classificada;

4.1 Coleta de dados

* Como Twitter expõe seus dados * Começando na API, Como podemos usá-la (autenticação como dev) * Quantidade de requests por "janela"* Resource Search * Filtros hashtag, linguagem(falar sobre problema), posição geográfica, etc * Exemplo do retorno *

A plataforma do Twitter conecta aplicações e sites com seus dados através de diversos serviços. Para este trabalho, nossa principal fonte de dados será sua API REST, que

possui uma excelente documentação disponível em [15]. Através dela é possível acessar informações de usuários e *tweets*, assim como escrever novas mensagens. Além disso, a API conta com um mecanismo de busca poderoso, que será fundamental para a coleta de dados. Os dados são entregues no formato *Javascript Object Notation* (JSON).

4.1.1 Autenticando na API

Para que ter acesso à API antes é necessário possuir uma conta no Twitter e utilizar o protocolo de autenticação OAuth[29], cujo principal objetivo é permitir que uma aplicação se autentique em outra "em nome de um usuário". A aplicação pede permissão de acesso ao usuário, que possui a escolha de conceder permissão ou não. Um ponto importante: o usuário não precisa informar a sua senha para se autenticar, portanto a permissão continua vigente caso a senha do usuário se altere, o que permite que a aplicação não precise de manutenção caso isso aconteça, tornando-a mais resiliente. A autenticação por meio do OAuth necessita de três passos:

1. Aplicação cliente obtém chave de autenticação;
2. Usuário autoriza aplicação cliente na aplicação servidora;
3. Aplicação cliente troca a chave de autenticação pela chave de acesso;

4.2 Armazenamento

4.3 Classificação

- Aplicar técnicas de normalização no texto. As mesmas devem ser específicas para a língua portuguesa;
- Construir base de palavras e termos classificados utilizadas como insumo para o modelo matemático;
- Preparar uma massa de treino para validar o modelo matemático antes da execução;

4.4 Normalização do texto

A composição de um *tweet* escrito por muitas vezes possui elementos que serão inúteis ou nocivos para o nosso algoritmo de classificação. Por conta disso, um dos primeiros

desafios para tal é conduzir uma normalização nas mensagens, que serão nosso objeto de estudo.

4.4.1 Construção da base de palavras e termos

A construção da base de dados foi feita com o intuito de melhor expressar um sentimento de uma palavra ou texto, para a utilização do algoritmo. Para isso a base foi dividida em dois arquivos, positivos e negativos. Além dessa divisão foi utilizada outas bases criadas como: Re-li(referencia), SentiLex-PT [30], base da puc [31], emoticons [32]. Todas usando a língua portuguesa ou um linguajar universal, no caso dos emoticons e já estarem polarizadas. Essas bases têm em comum é serem feitas apenas de palavras, então ficou-se a dúvida de como a classificação funcionaria posteriormente quando aplicadas a um texto que as palavras podem não estar no mesmo contexto. Ex: "O flamengo jogou muito mal, mas fico feliz pela vitória", onde tem a palavra mal que já dá um tom negativo a frase , porém ao terminar de ler a frase encontrasse as palavras feliz e vitória que tem um contexto positivo. Com essas bases já citadas foi compreendida a necessidade de uma base mais específica para o linguajar utilizado na internet, constituído de gírias, abreviação e até erros de português, para isso foi criada uma base utilizando dados pegos do twitter a partir da marcação hashtagoscar2016.

4.4.2 Massa de treino

4.4.3 Massa de teste

4.4.4 Algoritmo

4.5 Plataforma de análise

Capítulo 5

Resultados e análises

Neste capítulo serão apresentados resultados obtidos e os processos necessários para a obtenção dos mesmos. Como visto anteriormente no Capitulo 2, uma das etapas necessárias para a Análise de Sentimento é a classificação de polaridade dos *tweets*. Durante a classificação, que gera o resultado da execução do Naive Bayes, foram utilizadas diversas formas de execução que serão apresentadas durante este capítulo.'

5.1 Cenários e parâmetros de teste

Durante a execução dos testes para a análise de resultados o ambiente utilizado foi:

- Sistema operacional: Linux Ubuntu
- Processador: Core i7
- Memória: 8GB
- Quantidade de *tweets*: 141798

5.2 Experimentos realizados e resultados

O primeiro teste realizado para a classificação da base obteve o seguinte resultado:

Analisando a tabela 5.1 é visto quais as bases utilizadas, nesse caso, Reli , PUC e Sentilex, as técnicas utilizadas nesse teste, *Stopwords* e *Stemming*, e o resultado que de 141798 *tweets*, 17350 foram positivos, 15517 negativos e 108931 neutros, levando 311.673 segundos para executar o teste.

Trazer tabela 1 pra ca

Adiciona ima-

Tabela 5.1: 1º teste

1º Teste	
Bases usadas	Tecnicas usadas
Sentilex PUC ReLi	Stopwords Stemming
Resultado	
Positivo	17350
Negativo	15517
Neutro	108931
Tempo	311.673 segundos

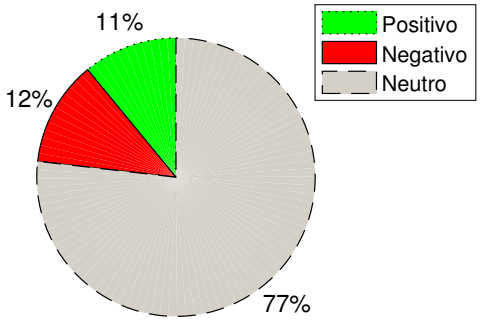

Figura 5.1: Quantidade de tweets separados por polaridade do teste 1. Fonte: Própria

Nota-se que a quantidade de *tweets* neutros é muito alta, evidenciando que o modelo ainda tem dificuldade de definir a polaridade do texto. Com base nos resultados apresentados foram realizadas as seguintes mudanças visando diminuir a ocorrência de "neutros".

Tabela 2 aqui

No 2º teste visto na tabela 5.2 é visto que a quantidade de neutro diminuiu consideravelmente, apenas aplicando a técnica de *stemming* nas bases de palavras

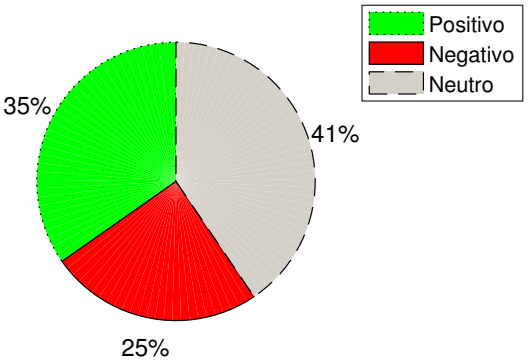

Figura 5.2: Quantidade de tweets separados por polaridade do teste 2. Fonte: Própria

Ainda buscando a diminuição de neutros foi criada uma base de palavras mas próxima do domínio que esse trabalho propõe que é o Oscar2016, essa base contem palavras relevantes a esse evento, gerando o seguinte resultado.

colocar tabela 3 aqui

Analisando a tabela 5.3 é visto que apenas uma base mais especializada no domínio não consegue diminuir a quantidade de neutros e ainda aumenta o tempo para a execução dos testes.

grafico 3

Tabela 5.2: 2º teste

2º Teste	
Bases usadas	Tecnicas usadas
Sentilex-Stem PUC-Stem ReLi	Stopwords-Stem Stemming
Resultado	
Positivo	49263
Negativo	35079
Neutro	57456
Tempo	397.48 segundos

Tabela 5.3: 3º teste

3º Teste	
Bases usadas	Técnicas usadas
Oscar2016	*Stopwords*
Resultado	
Positivo	47450
Negativo	7210
Neutro	87138
Tempo	709,126 segundos

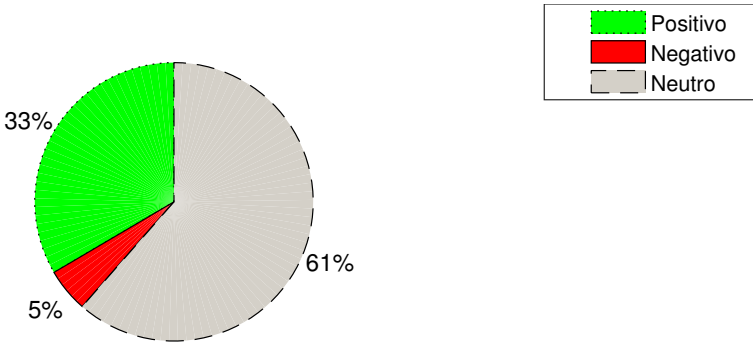

Figura 5.3: Quantidade de tweets separados por polaridade do teste 3. Fonte: Própria

No 4º teste foi adicionada o base criada, Oscar2016, com as bases genéricas, gerando o seguinte resultado:

Analisando a tabela 5.4 é visto que nesse teste foi obtida a maior diminuição de neutros, mas com um tempo de processamento um pouco maior. Segue abaixo um comparativo dos testes.

Adiciona gra-fico 4

Tabela 5.4: 4º teste

4º Teste	
Bases usadas	Técnicas usadas
Oscar2016 SentiLex PUC ReLi	Stopwords Stemming
Resultado	
Positivo	69070
Negativo	33461
Neutro	39267
Tempo	650.97 segundos

Tabela 5.5: Comparando testes

	Teste 1	Teste 2	Teste 3	Teste 4
Positivo	15517	49263	47450	69070
Negativo	17350	35079	7210	33461
Neutro	108931	57456	87138	39267
Tempo (s)	311.673	397.48	709.129	650.97

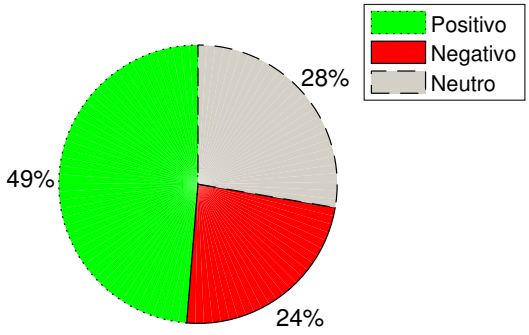

Figura 5.4: Quantidade de tweets separados por polaridade do teste 4. Fonte: Própria

Capítulo 6

Conclusão

Um paragrágo relembrando a importancia do cenário

Esse trabalho identificou e abordou alguns desses problemas, assim como propôs, desenvolveu e avaliou um serviço de gerenciamento eXXXXX Relembrar o que o trabalho fez.

A proposta, XXX, se destacou pelo XXXX que apresentou quando comparada XXXX.

A proposta atingiu os seguintes objetivos, exemplo:

- permitiu que sejam usados IEDs mais simples pois a solução não precisa ser implementada nesses dispositivos;
- reduziu o tempo de convergência dos algoritmos, o atraso na entrega de dados e o tráfego na rede;
- atendeu aos requisitos da Norma IEC 61850;
- implementou e testou um encaminhamento *multicast* independente de camadas e transparente aos dispositivos finais;
- permitiu uma configuração da rede facilitada;
- usou o arquivo SCD da norma para autoconfiguração da rede de Telecomunicações;
- tornou a rede menos sujeita à erros por ser automático;
- permitiu o uso mais inteligente de recuperação de falhas;
- permitiu o alcance de tempos de resposta menores por possuir uma característica proativa.

Os experimentos e as análises realizadas mostraramXXXXXX

Falar de todos os resultados encontrados de forma sumarizada, máximo de uma folha.

Os testes mostraram, também, que

Outro ganho relacionado ao uso da técnica....

A análise realizada mostra que ...

6.1 Trabalhos Futuros

Como trabalhos futuros, pretende-se ...

Uma outra questão é o estudo, desenvolvimento e implementação ...

Por fim, pretende-se fazer ...

Referências

[1] H. M. M. Lastres, S. Albagli, and C. A. K. Passos, *Informação e globalização na era do conhecimento*. Campus Rio de Janeiro, 1999.

[2] D. Santos, "O projecto processamento computacional do português: Balanço e perspectivas," *quot; In Maria das Graças Volpe Nunes (ed) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)(Atibaia SP 19-22 de Novembro de 2000) São Paulo: ICMC/USP*, 2000.

[3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[4] K. R. Scherer and M. R. Zentner, "Emotional effects of music: Production rules," *Music and emotion: Theory and research*, pp. 361–392, 2001.

[5] F. L. d. Santos, "Mineração de opinião em textos opinativos utilizando algoritmos de classificação," 2014.

[6] C. A. S. R. et al., "Mineração de opinião / análise de sentimento." [Online]. Available: http://www.inf.ufsc.br/~alvares/INE5644/MineracaoOpiniao.pdf

[7] G. Villela and P. A. Mendes, "Finanças comportamentais: O impacto da razão e da emoção no processo decisório em investimentos no mercado financeiro brasileiro," *Revista de Administração da FATEA*, vol. 6, no. 6, pp. 81–92, 2013.

[8] [Online]. Available: "http://www.whatdoestheinternetthink.net"

[9] (2016, Maio) Twitter company. [Online]. Available: https://about.twitter.com/company

[10] (2016, Abril) Dmr stats. [Online]. Available: http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/

[11] (2016, Maio) Twitter company. [Online]. Available: "https://dev.twitter.com/overview/api/counting-characters"

[12] M. Waite, *Paperback Oxford English dictionary*. Oxford University Press, 2012.

[13] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Mazaid, "Opening closed regimes: what was the role of social media during the arab spring?" *Available at SSRN 2595096*, 2011.

[14] S. Wasserman and J. Galaskiewicz, *Advances in social network analysis: Research in the social and behavioral sciences*. Sage Publications, 1994, vol. 171.

[15] (2016, Maio) Twitter company. [Online]. Available: "https://dev.twitter.com/overview/documentation"

[16] R. Fielding, "Architectural styles and the design of network-based software architectures." [Online]. Available: http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm

[17] M. A. Covington, *Natural language processing for Prolog programmers*. Prentice Hall Englewood Cliffs (NJ), 1994.

[18] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[19] J. D. Rennie, L. Shih, J. Teevan, D. R. Karger *et al.*, "Tackling the poor assumptions of naive bayes text classifiers," in *ICML*, vol. 3. Washington DC), 2003, pp. 616–623.

[20] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering," *arXiv preprint cs/0006013*, 2000.

[21] G. Lucca, I. A. Pereira, A. Prisco, and E. N. Borges, "Uma implementação do algoritmo naïve bayes para classificação de texto," 2013.

[22] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[23] H. J. C. Gomes, "Text mining: análise de sentimentos na classificação de notícias," Ph.D. dissertation, 2013.

[24] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREc*, vol. 10, 2010, pp. 1320–1326.

[25] P. L. Tortella and J. M. A. Coello, "Análise de sentimentos em mídias sociais."

[26] G. A. Rodrigues Barbosa, I. S. Silva, M. Zaki, W. Meira Jr, R. O. Prates, and A. Veloso, "Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment," in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2012, pp. 2621–2626.

[27] T. Franca and J. Oliveira, "Análise de sentimento de tweets relacionados aos protestos que ocorreram no brasil entre junho e agosto de 2013," in *III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2014.

[28] J. A. CARVALHO FILHO, "Mineração de textos: Análise de sentimento utilizando tweets referentes à copa do mundo 2014," 2014.

[29] [Online]. Available: "http://oauth.net/"

[30] P. C. "MÃ¡rio J. Silva and L. Sarmento", ""lecture notes in computer science"," in *"Building a Sentiment Lexicon for Social Judgement Mining"*, "International Conference on Computational Processing of the Portuguese Language (PROPOR)". "Springer", "2012", pp. "218–228".

[31] C. Freitas, “Sobre a construção de um léxico da afetividade para o processamento computacional do português,” *Revista Brasileira de Linguística Aplicada*, vol. 13, no. 4, pp. 1013–1059, 2013.

[32] F. F. M. B. F. d. J. "Alexander Hogenboom, Daniella Bal and U. Kaymak". "emoticon sentiment lexicon". [Online]. Available: "http://people.few.eur.nl/hogenboom/files/EmoticonSentimentLexicon.zip"