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

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]

Com a crescente popularidade de blogs e redes sociais, as áreas de mineração de opinião e análise de sentimento se tornaram objeto de estudo de pesquisadores. Uma abordagem ampla sobre o assunto foi apresentada em Pang e Lee [3]. 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 [4] a mineração de texto é aplicada em busca de notícias sobre economia em 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 [5] 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 como objeto de estudo, como por exemplo Tortella e Coelho [6]. Outros, se propõem a estudar um evento ou acontecimento finito, como por exemplo as eleições presidenciais no Brasil no ano de 2010 [7], os protestos populares contra a corrupção ocorridos em 2013 [8] e a Copa do Mundo da FIFA Brasil 2014 [9]. 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 a fim de determinar a relevância, impacto e opiniões geral 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 [5], porém utilizando português 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 [7] [8] [9].

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 3 , 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 3, onde a arquitetura da proposta é detalhada, assim como seus componentes e algoritmos. Em seguida, o Capítulo 4 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 5 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, 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 há uma grande urgência para novas teorias computacionais e ferramentas que ajudem a analisar essa quantidade de dados gerados [10]. 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.

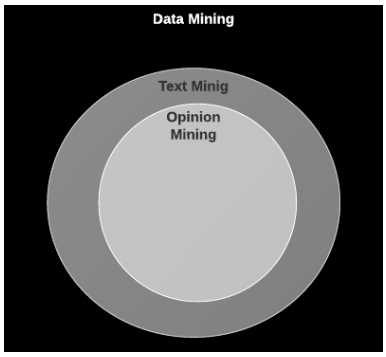

Figura 2.1: Diagrama de Venn - Mineração de Dados

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[11]. 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 [12] que tem como objetivo extrair o sentimento do texto escrito por uma pessoa, sem a interferência humana durante o processo.

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: [13]

- 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* - uma rotina sistemática capaz de varrer sites em busca de informações - 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 em positivo, negativo e neutro.

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 3.

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 empresa 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 [14]. 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[15], 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 [16], o Twitter é conhecido como um *microblog* fundado em março de 2006 por Jack Dorsey, Evan Williams e Biz Stone. Os usuários trocam mensagens de até 140 caracteres [17] em um ambiente de rede social, que tem como objetivo dar à todos o poder de compartilhar ideias e informações instantaneamente [16]. 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* [18] - nome pelo qual as mensagens compartilhadas no microblog ficaram conhecidas na Internet - por dia.

Dentro do Twitter, O usuário pode fazer uso de *hashtags* - marcadores conhecidos do público de rede social, que servem como uma indexação para um tópico específico [19]. 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 [20]. 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

A análise de redes sociais ganhou incrível relevância nos campos de pesquisa social e comportamental[21] com o aumento de compartilhamento de informações dentro destes ambientes. 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 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 - conjuntos de duas, três ou mais atores sociais. 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. Por exemplo, os dados extraídos a partir da API do Twitter, tema 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

Uma API é um conjunto de rotinas estabelecidos por um software para a utilização de suas funcionalidades e acessos aos seus dados por outro software que não pretende fazer uso de sua implementação, apenas de 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 simples.

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* e *mobile*. 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.

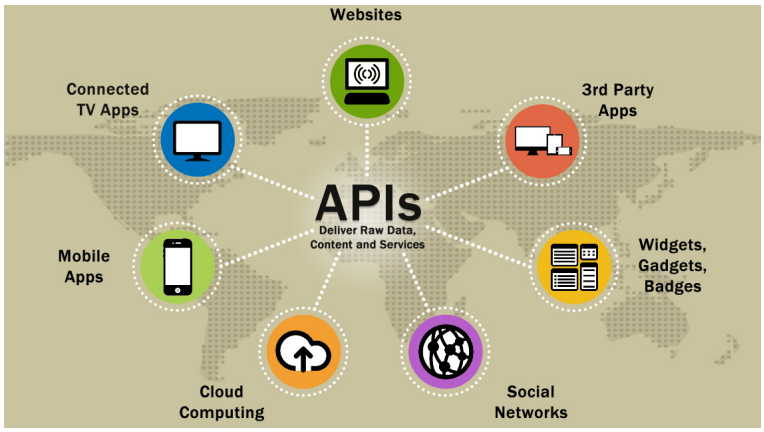

Figura 2.3: Papel das APIs integrando dados e serviços em diferentes plataformas. Fonte: http://www.programmableweb.com/

O Twitter, nossa fonte de dados durante este trabalho, possui uma API pública que pode ser utilizada por qualquer usuário da rede social [22].

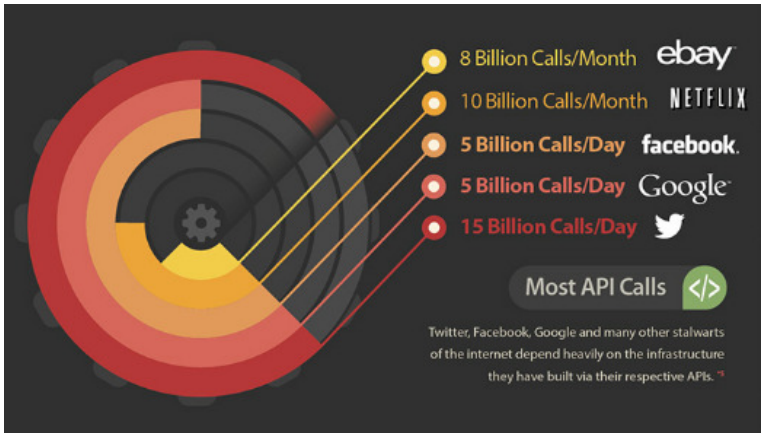

Figura 2.4: APIs mais utilizadas do mundo Fonte: SmartFile

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

2.3.1 REST

O protocolo REST foi criado em 2000 por Roy Fielding [23] como parte de sua dissertação de doutourado na *University of California Irvine*. Suas principais vantagens são:

- Por ter sido criado dentro de um ambiente acadêmico, o objetivo do protocolo abraça a filosofia *open source* - que preza por projetos onde o código é aberto a todos para manutenção e colaboração [24];
- 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 prevenindo 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, as APIs do Twitter, LinkedIn e Slack.

2.3.2 SOAP

Criado em 1998 por Dave Winer et al com colaboração da Microsoft, o protocolo SOAP concentra-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. [25].

A pesquisa na área está voltada a quatro 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, o PNL será 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 teste fundamenta conceitos chave sobre a Inteligência Artificial, 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* [26], 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 apresenta 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.

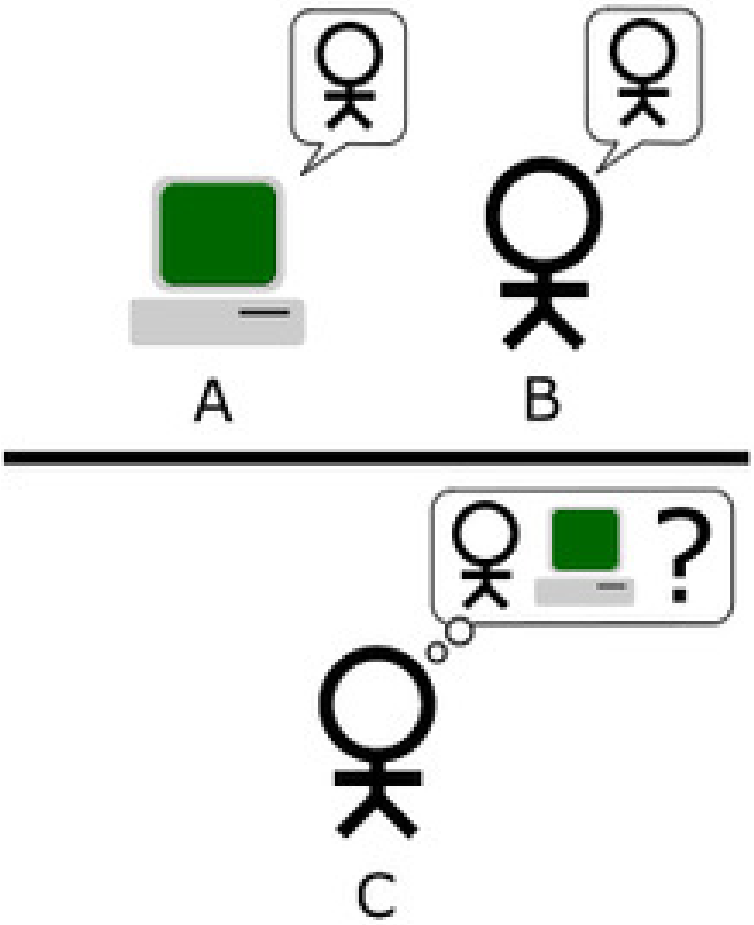

Figura 2.5: O participante A (máquina) e o participante B (humano) se comunicam por texto com o participante C (juiz). Fonte: Wikipédia

2.5 Classificador Naive Bayes

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

O classificador conhecido como *Naive Bayes* é um algoritmo probabilístico baseado no Teorema de Bayes que não considera que eventuais dependências possam existir. Por este motivo, suas suposições são nomeadas "ingênuas- de onde surge o nome *naive* - o que lhe confere uma maior simplicidade e um desempenho maior, em relação a outros algoritmos de classificação [27]. É um método popular para categorização de textos, como por exemplo a classificação de *e-mails* em legítimos ou *spam* - e-mails inúteis que são enviados na esperança que o receptor compre algum produto ou serviço. [28]

2.5.1 O Teorema de Bayes

O Teorema de Bayes permite inferir 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.

- 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 o evento B ocorreu.
- P(B|A) é a probabilidade de observar o evento B, dado que o evento A ocorreu.

Suponha que queremos saber a probabilidade de um indivíduo possuir 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, P(A). 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 ou P(B). 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 P(A|B).

$$(\frac{0,5}{100} \times \frac{1}{100}) \div \frac{0,2}{100} = \frac{2,5}{100}$$

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

2.5.2 Aplicação no trabalho

Nese trabalho o *Naive Bayes* é utilizado 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 [29] o algoritmo calcula qual a probabilidade de uma frase, denominada documento pertencer a uma determinada classe(polaridade) *P(C/D)*, a partir da probabilidade de *P(C)* do documento pertencer a esta classe e das probabilidades condicionais de cada termo t_k ocorrer em um documento da mesma classe. O algoritmo tem como objetivo encontrar a melhor classe para um documento maximizando a probabilidade conforme a equação abaixo, onde n_d é o número de termos no documento *d*.

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

Capítulo 3

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;

3.1 Coleta de dados

A plataforma do Twitter conecta aplicações e sites com seus dados através de diversos serviços. Para este trabalho, a principal fonte de dados será sua API REST, que possui uma excelente documentação disponível em [22]. 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).

3.1.1 Autenticação

Para que ter acesso à API antes é necessário possuir uma conta no Twitter e criar uma *app* - através do próprio site [30] - que utilizará o protocolo de autenticação OAuth[31] para acessar os dados do Twitter se passando pelo usuário em questão. O objetivo do protocolo OAuth 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 neste caso, 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;

Após o processo de criação da app, é criado um token de acesso que deve ser utilizado pelo sistema que deseja se autenticar no Twitter em nome de um usuário. Este token deve ser incorporado em cada requisição à API do Twitter para autenticar a mesma e dizer ao Twitter qual é a fonte do acesso.

3.1.2 Limite de requisições

A fim de evitar grande concentração de requisições em seus serviços, o Twitter implementa um limitador em sua API [32]. São permitidas até 180 requisições por janela, que dura 15 minutos. Caso o limite seja ultrapassado, o serviço passa a retornar um erro na resposta, até que a "janela"de 15 minutos se renove. A partir da versão 1.1 da API, novos cabeçalhos HTTP são retornados provendo feedback sobre os limites para requisição. Este recurso permite que o código consiga entender em que momento da janela se encontra, quantos requisições ainda podem ser feitas neste período de tempo e quanto é necessário esperar para poder fazer novas requisições. Os cabeçalhos em questão são:

- X-Rate-Limit-Limit: A faixa limite para o requisição em questão;

- X-Rate-Limit-Remaining: O número de requisições que ainda restam para a janela de 15 minutos;
- X-Rate-Limit-Reset: O tempo restante dentro da janela de requisições atual, dado em segundos.

3.1.3 Arquitetura

Neste trabalho, como o objetivo é coletar *tweets* postados sobre uma *hashtag* em tempo real para utilizá-los como matéria-prima para análise de sentimento, é muito importante aproveitar ao máximo cada janela de requisições. Por este motivo, o sistema que coleta os dados da API do Twitter foi inspirado no modelo produtor-consumidor[33] visando minimizar as perdas que podem acontecer em momentos de pico - como o começo ou clímax do evento, onde o volume de mensagens é maior, como veremos a frente no Capítulo 4 - e se necessário, escalar de forma simples durante os mesmos.

3.1.3.1 Produtor-consumidor

O problema descreve dois processos, o produtor e o consumidor, que compartilham um recurso em comum usado como uma fila - um tipo particular de coleção de dados onde a primeira a entidade a entrar é a primeira a sair ou *First-In-First-Out* (FIFO). A função do produtor é gerar trabalho a ser executado pelo consumidor. O volume de trabalho gerado e executado pelo sistema é controlado pela fila, que armazena as entidades ou "tarefas"a serem executadas. Essa abordagem permite que o sistema escale apenas até a sua capacidade, visto que a fila possui um tamanho fixo que caso seja ultrapassado, pode simplesmente descartar as mensagens adicionadas após este momento. Outra característica importante é a escalabilidade. Conforme os processos produtor e consumidor evoluem, surge a necessidade de aumentar a quantidade de produtores ou consumidores de forma independente.

Neste trabalho, para explorar o potencial máximo da janela de requisições foi criado um processo produtor que envia para a fila mensagens para que consumidor acesse à API do Twitter de forma que sejam feitos sempre as 180 requisições que são permitidas no intervalo de 15 minutos. Assim a responsabilidade de cada processo fica bem definida - o primeiro responde pelo volume de requisições e o segundo por realizar a requisição e entender a resposta. Para definir qual intervalo de tempo deveria ser usado para que o produtor envie mensagens à fila, foi feita uma conta simples:

$$\frac{180}{15} = 12 \ requests_{/minuto}$$

Logo, o produtor precisa adicionar uma mensagem na fila a cada 5 segundos, para que o limite de 180 requisições seja respeitado.

Fazer diagrama do produtor consumi para explicar melhor

3.1.4 Busca

Dentre os principais serviços da API do Twitter está a busca. Com ela, é possível consultar de diversas formas os principais *tweets* ou mais recentes. Dentro de sua documentação, existe um guia completo de como utilizar a API para extrair os resultados desejados [34] das mais diversas formas.

3.1.4.1 Parâmetros adicionais

Como abordado acima, a API possui diversos parâmetros que podem ser usados para que o usuário chegue a um conjunto de dados mais próximo da sua necessidade:

- ***result_type***: permite escolher se o resultado da busca será representado pelos *tweets* mais populares (*popular*) ou mais recentes (*recent*);
- ***geocode***: permite buscar por uma determinada latitude, longitude e raio, respectivamente, separando-os por vírgula. ex: geocode=-22.912214,-43.230182,1km;
- ***lang***: restringe os *tweets* buscados a um idioma específico. ex: lang=pt;
- ***since_id*** , ***max_id*** , ***count*** e ***until***: possibilita iterar através dos resultados quando existe um grande números de *tweets* a percorrer. De acordo com a concorrência e o volume, esta tarefa pode ficar mais complicada. Uma leitura recomendada se encontra em [35].

Como o objetivo deste trabalho é coletar novos *tweets* conforme eles vão sendo postados, foi necessário utilizar apenas dois parâmetros da API de busca: *count* e *since id*. O primeiro tem o objetivo de garantir o número máximo de registros retornados pela API e o segundo define de onde se pretende partir para buscar novos *tweets*, evitando que mensagens repetidas sejam coletadas.

`https://api.twitter.com/1.1/search/tweets.json?q=#oscars2016&count=100&since_id=123456789`

Neste caso, a API retornará os 100 *tweets* publicados desde o *tweet* com identificador (*id*) "123456789".

Diagrama de sequência ou atividades sobre como esta parte do código funciona

3.1.4.2 O problema com a detecção automática de idioma do Twitter

O escopo deste trabalho determina que o objeto de estudo são apenas mensagens escritas em língua portuguesa. Como forma de obter somente *tweets* escritos em língua portuguesa é preciso utilizar o parâmetro *lang*, por exemplo:

`https://api.twitter.com/1.1/search/tweets.json?q=#oscars2016&lang=pt&result_type=recent`

Porém, realizando alguns testes na API do Twitter, foi detectado que ao submeter algum *tweet*, alguma rotina dentro do próprio Twitter atribui um idioma à mensagem automaticamente. Nos testes conduzidos durante este trabalho foi identificado que em mensagens curtas - algo recorrente no Twitter - o algoritmo apresenta resultados aquém do esperado na identificação do idioma, visto que as poucas palavras contidas na mensagem podem ser comuns a mais de um idioma. Por conta disso foi decidido que o filtro de idioma não seria utilizado. Esta decisão pode mudar de acordo com o evento monitorado ou com o escopo do estudo.

3.1.4.3 Escalando de forma horizontal

Por uma questão de desenho da API de busca, cada requisição é capaz de trazer no máximo 100 *tweets*, o que nos dá ao total uma carga máxima de até 1200 novos *tweets* por minuto. Para as análises feitas durante este trabalho este número se mostrou mais do que suficiente. Como dito anteriormente, se fosse necessário escalar este sistema para monitorar um evento maior, seria necessário apenas obter um saldo maior de requisições junto a API do Twitter - adicionando mais tokens de usuário e criando uma espécie de "rodízio"de autenticações, por exemplo - e escalar o número de consumidores do processo de acordo com a demanda. Uma boa maneira de detectar se isso seria necessário é acompanhar quantos *tweets* novos são coletados a cada requisição. Como utilizamos o *id* do último *tweet* capturado como referência para os novos, se a cada requisição o número de novos *tweets* com grande frequência coincidir com o limite da API, temos um indício

de que o volume de novas mensagens no Twitter está excedendo a capacidade do sistema de coletá-las e que o excedente está sendo perdido.

3.2 Armazenamento

A resposta da API de Busca é dada no formato JSON e cada objeto - que corresponde a cada *tweet* - segue o seguinte formato e conta com diversas informações sobre o mesmo:


```
{
        "_id" : "56d388096861353c1b061fe2",
        "contributors" : null,
        "truncated" : false,
        "text" : "Leonardo DiCaprio com o #oscars eh igual a
           Katy Perry com o Grammy #OscarsNaTNT",
        "is_quote_status" : false,
        "in_reply_to_status_id" : null,
        "id" : 704091511902834688,
        "favorite_count" : 0,
        "source" : "<a href=\"http://twitter.com/download/
           android\" rel=\"nofollow\">Twitter for Android</a
           >",
        "created_at_datetime" : ISODate("2016-02-28T20:51:
           12.000Z"),
        "retweeted" : false,
        "coordinates" : null,
        "created_at_timestamp" : 1456714272.0000000000000000,
        "entities" : {
                "symbols" : [],
                "user_mentions" : [],
                "hashtags" : [{
                        "indices" : [ 24, 31 ],
                        "text" : "oscars"
                },
                {
                        "indices" : [ 66, 78],
                        "text" : "OscarsNaTNT"
```

```
                }],
                "urls" : []
        },
        "in_reply_to_screen_name" : null,
        "id_str" : "704091511902834688",
        "retweet_count" : 0,
        "in_reply_to_user_id" : null,
        "favorited" : false,
        "user" : {
                "follow_request_sent" : false,
                "has_extended_profile" : true,
                "profile_use_background_image" : false,
                "id" : 2786117482,
                "verified" : false,
                "profile_text_color" : "000000",
                "profile_image_url_https" : "https://
                  pbs.twimg.com/profile_images/
                  700450356141158404/xA-mRqp7_normal.jpg",
                "profile_sidebar_fill_color" : "000000",
                "is_translator" : false,
                "entities" : {
                        "description" : {
                        "urls" : []
                        }
                },
                "followers_count" : 156,
                "protected" : false,
                "location" : "Um lugarzinho no fim do mundo",
                "default_profile_image" : false,
                "id_str" : "2786117482",
                "lang" : "pt",
                "utc_offset" : -28800,
                "statuses_count" : 6171,
                "description" : "Uuuumm ta estilosa!",
                "friends_count" : 192,
```

```
58                "profile_background_image_url_https" : "https
                    ://abs.twimg.com/images/themes/theme1/
                    bg.png",
59                "profile_link_color" : "9266CC",
60                "profile_image_url" : "http://pbs.twimg.com/
                    profile_images/700450356141158404/xA-
                    mRqp7_normal.jpg",
61                "notifications" : false,
62                "geo_enabled" : false,
63                "profile_background_color" : "000000",
64                "profile_banner_url" : "https://pbs.twimg.com
                    /profile_banners/2786117482/1456444936",
65                "profile_background_image_url" : "http://
                    abs.twimg.com/images/themes/theme1/bg.png"
                    ,
66                "name" : "Padeira Estilosa",
67                "is_translation_enabled" : false,
68                "profile_background_tile" : false,
69                "favourites_count" : 6095,
70                "screen_name" : "naycordeir",
71                "url" : null,
72                "created_at" : "Fri Sep 26 18:43:47 +0000
                    2014",
73                "contributors_enabled" : false,
74                "time_zone" : "Pacific Time (US & Canada)",
75                "profile_sidebar_border_color" : "000000",
76                "default_profile" : false,
77                "following" : false,
78                "listed_count" : 1
79           },
80           "geo" : null,
81           "in_reply_to_user_id_str" : null,
82           "lang" : "pt",
83           "created_at" : "Sun Feb 28 23:51:12 +0000 2016",
84           "metadata" : {
```

```
85                  "iso_language_code" : "pt",
86                  "result_type" : "recent"
87          },
88          "in_reply_to_status_id_str" : null,
89          "place" : null
90  }
```

Dentro desta resposta, existem diversos dados que podem ser úteis para análises em cima dos *tweets* e armazená-los é extremamente valioso.

3.2.1 Banco de dados orientado a documento

Existem diversas soluções de banco de dados disponíveis no mercado. Nos últimos anos, uma delas se tornou especialmente popular[36]: os bancos de dados orientado a documento. Tais bancos de dado são uma das principais categorias de bancos conhecidos NoSQL (*Non Structure Query Language*) que consiste em organizar os dados de forma "não-relacional", através de documentos, gráficos, chave-valores e colunas. Bancos NoSQL são conhecidos pela facilidade de modelagem e desenvolvimento, alto desempenho de leitura e escrita, alta disponibilidade e resiliência. Isso não significa que bancos SQL são obsoletos ou piores, porém existem aplicações claras onde cada um desempenha um melhor papel. Podemos apontar algumas comparações, como por exemplo:

	Banco de dados relacional	Banco de dados NoSQL
Modelagem	O modelo relacional normaliza dados em estruturas tabulares conhecidas como tabelas, que consistem em linhas e colunas. Um schema define estritamente as tabelas, colunas, índices, relações entre tabelas e outros elementos do banco de dados.	Bancos de dados não relacionais (NoSQL) normalmente não aplicam um schema. Geralmente, uma chave de partição é usada para recuperar valores, conjuntos de colunas ou documentos semiestruturados JSON, XML ou outros que contenham atributos de itens relacionados.
Desempenho	O desempenho normalmente depende do subsistema do disco. A otimização de consultas, índices e estrutura de tabela é necessária para alcançar máximo desempenho.	Desempenho geralmente é uma função do tamanho do cluster do hardware subjacente, da latência de rede e da aplicação que faz a chamada.
Escala	Mais fácil de aumentar a escala "verticalmente"com hardware mais rápido.,Outros investimentos são necessários para tabelas relacionais para abranger um sistema distribuído.	Projetado para aumentar a escala "horizontalmente"usando clusters distribuídos de hardware de baixo custo para aumentar a transferência sem aumentar a latência.
APIs	As solicitações para armazenar e recuperar dados são comunicadas usando consultas compatíveis com structured query language (SQL). Essas consultas são analisadas e executadas por sistemas de gerenciamento de bancos de dados relacionais (RDBMS).	APIs baseadas em objetos permitem que desenvolvedores de aplicações armazenem e restaurem facilmente estruturas de dados na memória. As chaves de partição permitem que os aplicativos procurem pares de chave-valor, conjuntos de colunas ou documentos semiestruturados contendo objetos e atributos de aplicativos serializados.

Tabela 3.1: Comparação entre bancos SQL e NoSQL

3.2.2 Opções disponíveis

O movimento de adoção de bancos *NoSQL* está bastante enraizada no mundo *open source*. Alguns projetos como Voldemort[37], MongoDB[38], Tokyo Cabinet[39] e CouchDB[40]. Apesar de uma grande quantidade de opções *open source*, o movimento ganhou muita força com a publicação de duas publicações sobre implementações proprietárias: o Google BigTable[41] e o Amazon Dynamo[42]. Para este trabalho, a opção escolhida foi o MongoDB, altamente popular na comunidade *open source* e com bastante material disponível com melhores práticas de criação, manutenção e configuração.

3.2.3 Armazenando *tweets*

O objetivo deste trabalho é monitorar eventos através das *hashtags*. E para cada uma delas é criada uma coleção - nome dado a um conjunto de documentos - dentro do banco de dados. O nome da coleção é dado pela *hashtag* monitorada.

No momento onde o processo consumidor - responsável por fazer requisições na API e lidar com o retorno - é ligado ocorre a criação da coleção, caso a mesma não exista, e dentro dela começam a ser armazenados os *tweets* retornados pela API, obedecendo ao mesmo formato enviado pelo Twitter. A flexibilidade dAntes do armazenamento no banco alguns campos a mais são adicionados, mas vamos entrar neste detalhe apenas na seção sobre Classificação.

3.3 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;
- Calibragem do algoritmo
- Salvar infos sobre a classificação dentro do documento

3.3.1 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.

3.3.2 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 [43], base da puc [44], emoticons [45]. 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.

3.3.3 Massa de treino

3.3.4 Massa de teste

3.3.5 Algoritmo

Capítulo 4

Resultados e análises

Neste capítulo serão apresentados os processos e os resultados obtidos nesse trabalho. Como visto anteriormente no Capitulo 3, uma das etapas necessárias para a Análise de Sentimento é a classificação de polaridade dos *tweets*. Durante a classificação, resultado da execução do Naive Bayes, foram utilizados diferentes parâmetros que serão apresentados e discutidos durante este capítulo.

4.1 Cenários e parâmetros de teste

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

- Sistema operacional: Linux Ubuntu 15.04
- Processador: Core i7
- Memória: 8GB
- Quantidade de *tweets*: 141.798

4.2 Técnicas

Durante a realização do trabalho foram utilizados duas técnicas para maximizar a performance do modelo. As técnicas foram aplicadas nas bases utilizadas para a classificação e nos dados coletados de acordo com os testes discutidos ao longo do capítulo

Palavra	Stemização
boate	boat
boates	boat
boca	boc
bocados	boc

Tabela 4.1: Exemplo de stemização

4.2.1 Stemming

A técnica de *stemming*, conhecido em português como stemização, consiste na redução de um termo ao seu radical, removendo as desinências, afixos, e vogais temáticas. Com sua utilização, os termos derivados de um mesmo radical serão contabilizados como um único termo [46].

4.2.2 Stopwords

A técnica de *stopwords*, consiste na remoção de palavras "vazias", como artigos, preposições e interjeições, que não agregam valor a análise realizada. Além disso, essas palavras podem ser configuráveis dependendo do domínio do seu estudo [47].

4.3 Desenvolvimento do modelo de análise

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

1º Teste	
Bases usadas	Tecnicas usadas
Sentilex PUC ReLi	Stopwords Stemming
Resultado	
Positivo	17.350
Negativo	15.517
Neutro	108.931

Tabela 4.2: 1º teste

Como mostra a tabela 4.2 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

141.798 *tweets*, 17.350 foram positivos, 15.517 negativos e 108.931 neutros.

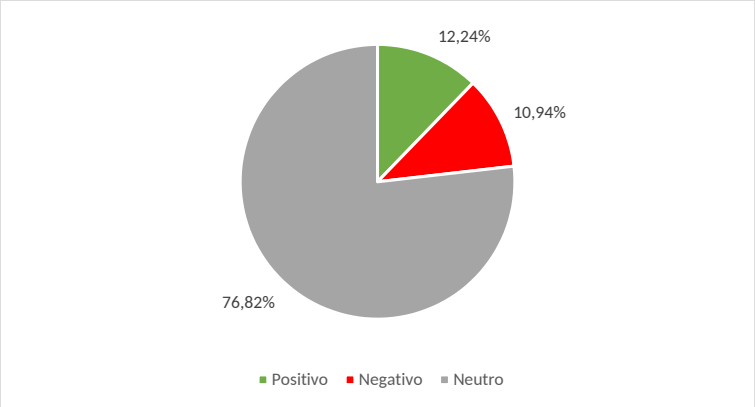

Figura 4.1: Quantidade de tweets separados por polaridade do teste 1. Fonte: Própria

Nota-se que a quantidade de *tweets* neutros é elevada, 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 dados neutros.

2º Teste	
Bases usadas	Tecnicas usadas
Sentilex-Stem PUC-Stem ReLi-Stem	Stopwords Stemming
Resultado	
Positivo	49.263
Negativo	35.079
Neutro	57.456

Tabela 4.3: 2º teste

Como mostra a tabela 4.3 as mesmas bases foram utilizadas, porém com a aplicação da técnica de *stemming*, como PUC-Stem e Sentilex-Stem. Além de utilizar a mesma técnica nas *stopwprds*. O resultado de 141.798 *tweets*, 49.263 foram positivos, 35.079 negativos e 57.456 neutros.

Analisando o 2º teste é visto que a quantidade de neutro diminuiu consideravelmente, apenas aplicando a técnica de *stemming* nas bases de palavras.

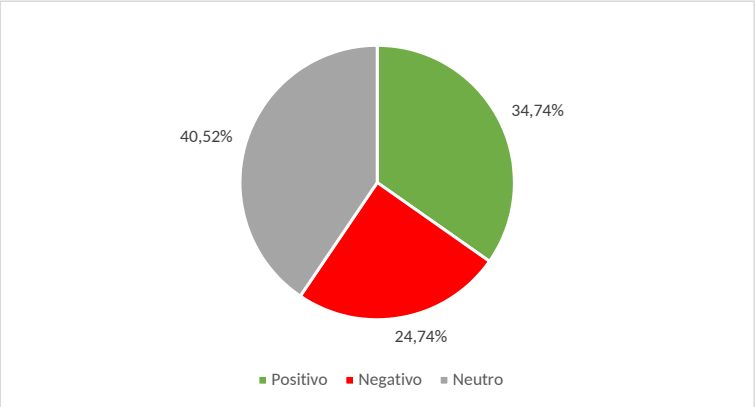

Figura 4.2: Quantidade de tweets separados por polaridade do teste 2. Fonte: Própria

Ainda buscando a diminuição de dados neutros foi criada uma base de palavras mais próxima do domínio que esse trabalho propõe com a base chamada Oscar2016, essa base contém palavras relevantes ao evento, gerando o seguinte resultado.

3º Teste	
Bases usadas	Técnicas usadas
Oscar2016	*Stopwords*
Resultado	
Positivo	47.450
Negativo	7.210
Neutro	87.138

Tabela 4.4: 3º teste

Analisando a tabela 4.4 é visto que apenas uma base mais especializada no domínio não consegue diminuir a quantidade de neutros. O resultado de 141.798 *tweets*, 47.450 foram positivos, 7.210 negativos e 87.138 neutros.

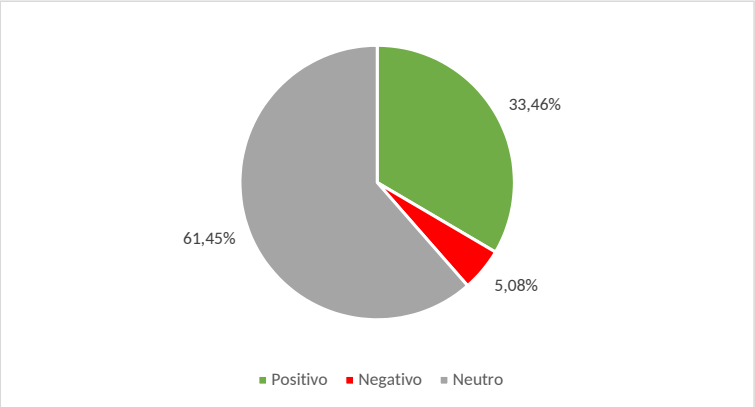

Figura 4.3: Quantidade de tweets separados por polaridade do teste 3. Fonte: Própria

No 4º teste foi adicionada a base criada, Oscar2016 com as bases genéricas, Sentilex, PUC e Reli gerando o seguinte resultado:

4º Teste	
Bases usadas	Técnicas usadas
Oscar2016 SentiLex PUC ReLi	Stopwords Stemming
Resultado	
Positivo	69.070
Negativo	33.461
Neutro	39.267

Tabela 4.5: 4º teste

Analisando a tabela 4.5 é visto que nesse teste foi obtido a maior taxa de diminuição de dados neutros. O resultado de 141.798 *tweets*, 69.070 foram positivos, 33.461 negativos e 39.267 neutros.

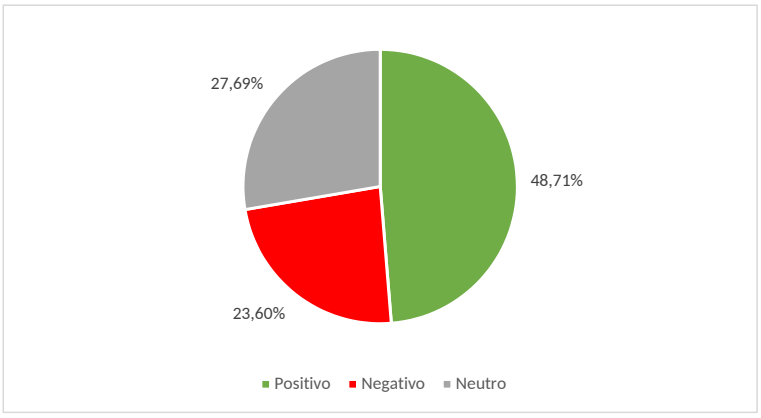

Figura 4.4: Quantidade de tweets separados por polaridade do teste 4. Fonte: Própria

Segue abaixo um comparativo dos testes.

	Teste 1	Teste 2	Teste 3	Teste 4
Positivo	15.517	49.263	47.450	69.070
Negativo	17.350	35.079	7210	33.461
Neutro	108.931	57.456	87.138	39.267

Tabela 4.6: Comparando testes

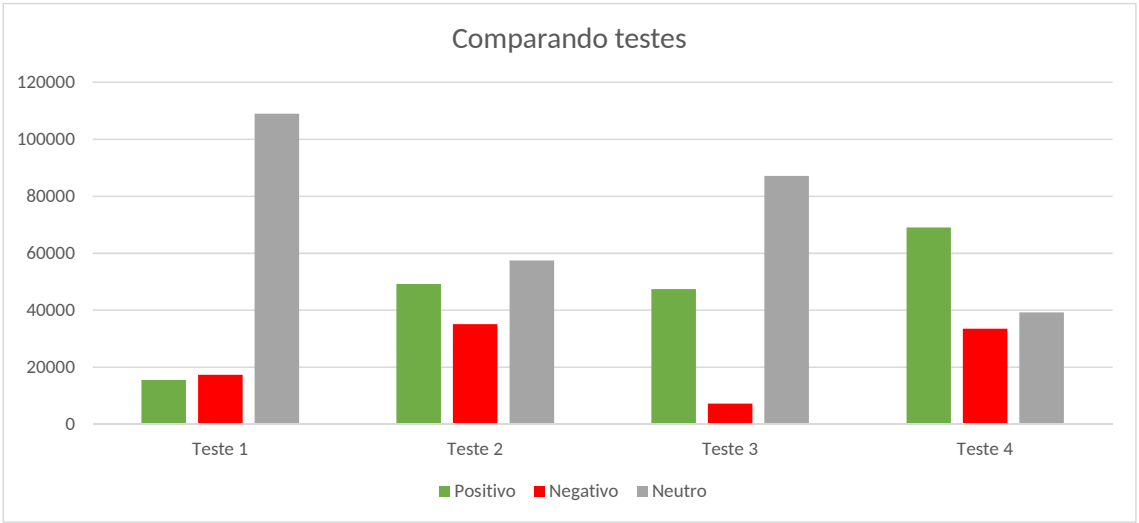

Figura 4.5: Gráfico de comparação dos testes

Com o comparativo analisado na tabela 4.6 foram escolhidas as configurações utilizadas no teste 4 para a realização da análise na base capturada, devido a menor presença de dados neutros classificados na base comprovando que as técnicas de *stemming* e *stopwords* foram válidas e relevantes para a sustentação modelo.

4.4 Resultados

Para a análise dos resultados é necessário estabelecer premissas. O evento Oscar 2016 foi a 88.ª cerimônia de entrega dos *Academy Awards* em *Los Angeles*, Estados Unidos. O evento teve duração de aproximadamente 4 horas com seu início na noite do dia 28 de fevereiro, as 20:00 horário de Brasília no tapete vermelho com a entrada dos artistas e convidados. Seu encerramento as 2:00 do horário de Brasília do dia 29 de Fevereiro com a premiação do melhor filme. Os marcos do evento foram considerados: as entregas dos prêmios de cada categoria, o homenagem em memória póstuma e a apresentação da Lady Gaga e o discurso do Vice Presidente Americano. Esses eventos são listados em ordem cronológica de acordo com a imagem 4.6

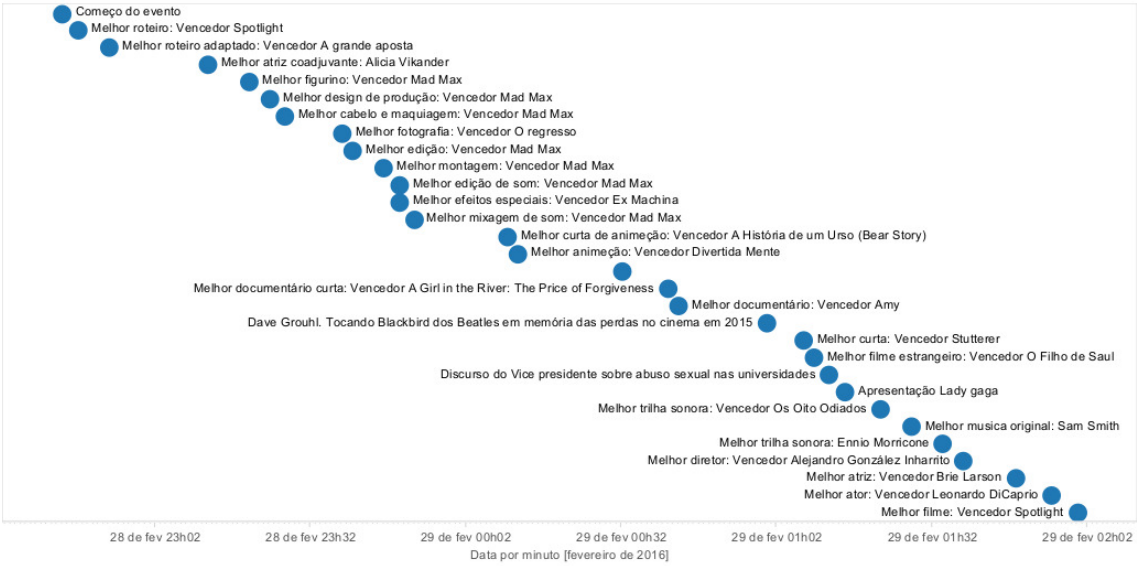

Figura 4.6: Linha do tempo com os marcos do Oscar 2016. Fonte:Folha de São Paulo

O gráfico 4.7 mostra a curva da quantidade de *tweets* em relação ao tempo dividido pela polaridade. Nele é visto um pico no começo do evento, durante o tapete vermelho, nesse período foi visto as maiores ocorrências de especulações sobre os ganhadores e indicados além das reações dos internautas com as entradas de seus artistas favoritos.

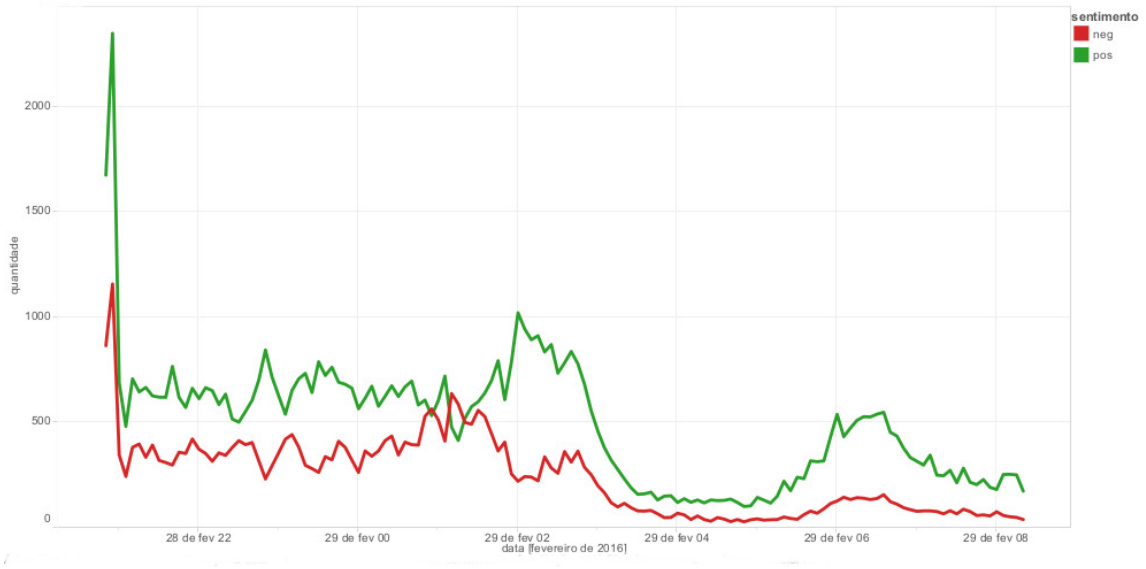

Figura 4.7: Quantidade de tweets por tempo e polaridade. Fonte:Própria

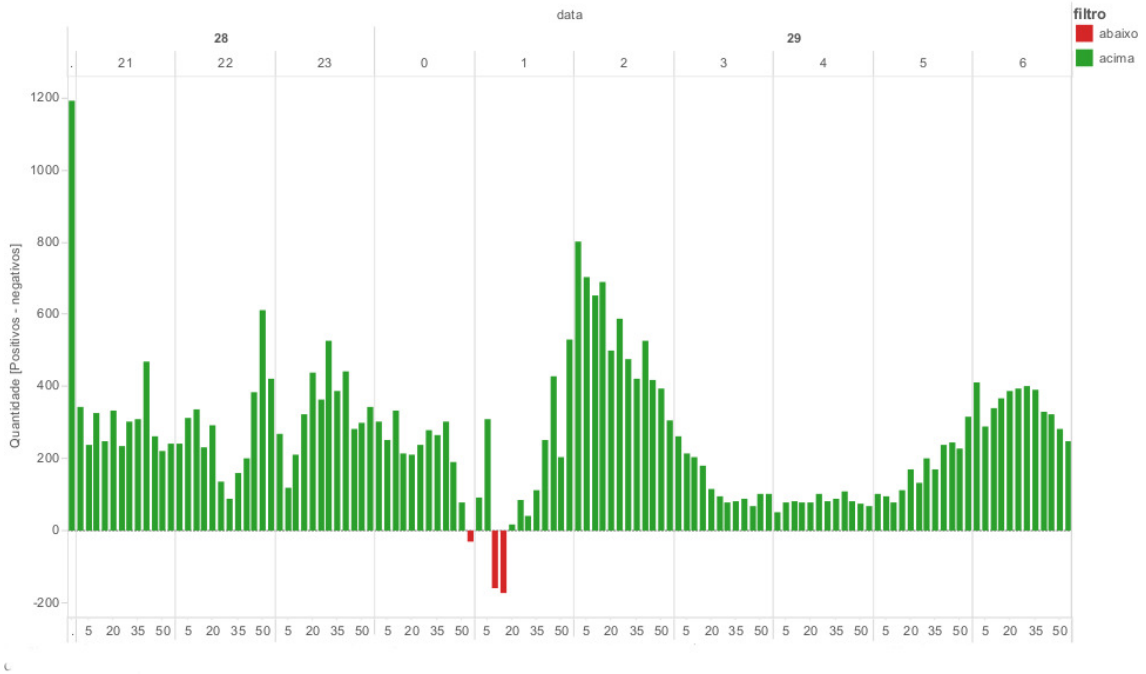

Figura 4.8: Quantidade de tweets positivos diminuído pelos negativos pelo tempo. Fonte:Própria

O gráfico 4.8 foi gerado a partir da diminuição dos *tweets* negativos pelos *tweets* positivos vistos no gráfico 4.7 produzindo uma visão micro do dados coletados do evento. Analisando o gráfico 4.7, que possui uma visão macro, junto com o gráfico 4.8, com a visão micro, é notado que em dois momentos a quantidade de *tweets* negativos sobrepõem a quantidade de positivos. No primeiro momento, é durante o *show* do Dave Grohl onde

é feita uma homenagem aos falecidos no mundo do cinema em 2015 onde a sensação de luto ficou evidente durante a apresentação, e de acordo com [48], é a reação à perda de objeto ou pessoa. Como afeto, o luto aproxima-se do humor depressivo. Com isso foi corroborado o aumento do sentimento negativo nessa parte do evento. Outro momento onde ocorre o mesmo efeito é durante a apresentação da Lady Gaga onde o tema do filme que a música faz parte da trilha é sobre os estupros em faculdades americanas, e mais uma vez esse aumento negativo faz entender o tema pesado e sensível para as pessoas que sofreram e sofrem disso. Excluindo o pico no começo dos gráficos os maiores picos de positivos são presenciados as 2:00, vindo de uma crescente no final das 1:00 da manhã do dia 29 isso se deve as categorias mais aguardadas do evento que são:

- Melhor atriz
- Melhor ator
- Melhor filme

O maior destaque foi a premiação de melhor ator e a vitória de Leonardo DiCaprio, que com seus 41 anos de idade e 26 de carreira, nunca havia levado um prêmio de melhor ator da Academia. Essa vitória reflete nos gráficos de picos de positivos as 2:00 da manhã onde a repercussão da sua vitória é espelhada pelo twitter chegando aos *trending topics* mundias na rede social.

Figura 4.9: Mapa de calor referente a polaridade de sentimento no Brasil. Fonte:Própria

Capítulo 5

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 ...

5.1 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] 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.

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

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

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

[7] 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.

[8] 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.

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

[10] 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.

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

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

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

[14] 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.

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

[16] (2016, Maio) Twitter company. [Online]. Available: https://about.twitter.com/company

[17] (2016, Maio) Twitter company. [Online]. Available: "https://dev.twitter.com/overview/api/counting-characters"

[18] (2016, Abril) Dmr stats. [Online]. Available: http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/

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

[20] 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.

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

[22] (2016, Maio) Twitter company. [Online]. Available: "https://dev.twitter.com/overview/documentation"

[23] 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

[24] S. Weber, *The success of open source.* Cambridge Univ Press, 2004, vol. 368.

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

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

[27] 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.

[28] 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.

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

[30] (2016, Maio) Twitter company. [Online]. Available: https://apps.twitter.com/

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

[32] (2016, Maio) Twitter company. [Online]. Available: "https://dev.twitter.com/rest/public/rate-limiting"

[33] K. Jeffay, "The real-time producer/consumer paradigm: A paradigm for the construction of efficient, predictable real-time systems," in *Proceedings of the 1993 ACM/-SIGAPP symposium on Applied computing: states of the art and practice*. ACM, 1993, pp. 796–804.

[34] (2016, Maio) Twitter company. [Online]. Available: "https://dev.twitter.com/rest/public/search"

[35] (2016, Maio) Twitter company. [Online]. Available: "https://dev.twitter.com/rest/public/timelines"

[36] N. T. Bhuvan and M. S. Elayidom, "A technical insight on the new generation databases: Nosql," *International Journal of Computer Applications*, vol. 121, no. 7, 2015.

[37] "voldemort project. [Online]. Available: "http://www.project-voldemort.com/voldemort/"

[38] "mongodb. [Online]. Available: "https://www.mongodb.com/"

[39] "tokyo cabinet: a modern implementation of dbm. [Online]. Available: "http://fallabs.com/tokyocabinet/"

[40] "couchdb. [Online]. Available: "https://couchdb.apache.org/"

[41] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," *ACM Transactions on Computer Systems (TOCS)*, vol. 26, no. 2, p. 4, 2008.

[42] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: amazon's highly available key-value store," *ACM SIGOPS Operating Systems Review*, vol. 41, no. 6, pp. 205–220, 2007.

[43] 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".

[44] 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.

[45] 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"

[46] V. M. Orengo and C. Huyck, "A stemming algorithm for the portuguese language," in *String Processing and Information Retrieval, 2001. SPIRE 2001. Proceedings.Eighth International Symposium on*, Nov 2001, pp. 186–193.

[47] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets:*. Cambridge University Press, Oct 2011. [Online]. Available: https://www.cambridge.org/core/books/mining-of-massive-datasets/A06D57FC616AE3FD10007D89E73F8B92

[48] S. Freud, "Conferências introdutórias sobre psicanálise," in *Conferência XXIII*, 1908.