

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.

Orientadora:
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. - Orientadora
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 comerciais, 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: xxxxxxxx.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

API	Application Program Interface	1
	Conjunto de rotinas estabelecidos por um software para a utilização das suas funcionalidades por aplicativos que não pretendem envolver-se em detalhes da implementação.	

Sumário

1	Introdução	1
1.1	Motivação e Objetivos	2
1.2	Principais contribuições	2
1.3	Recursos utilizados	2
1.4	Organização do trabalho	2
2	Referencial Teórico	3
2.1	Twitter	3
2.2	Mineração de opinião	3
2.3	API	4
2.4	Processamento de linguagem natural	4
2.5	Análise de sentimento	4
2.6	Naive Bayes	4
3	Proposta	5
3.1	Trabalhos relacionados	5
3.2	Implementação	5
3.2.1	Crawler	5
3.2.2	Classificação	5
3.2.2.1	Algoritmo	5
3.2.2.2	Construção da base de dados	5
3.2.2.3	Massa de treino	6

Sumário	x
3.2.2.4 Massa de teste	6
3.2.3 Plataforma de análise	6
4 Resultados e análises	7
4.1 Cenários e parâmetros de teste	7
4.2 Experimentos realizados e resultados	7
5 Conclusão	8
5.1 Trabalhos Futuros	9
Referências	10

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 à 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 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 Capítulo 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 Twitter

* Como começou * Objetivo (visão) do Twitter * Princípios 140 caracteres, hashtags * Quantidade de usuários ativos, alcance, volume de informações * Relevância para estudos estatísticos de natureza comportamental

O Twitter é conhecido como um *microblog* fundado em março de 2006 por Jack Dorsey, Evan Williams e Biz Stone. Ele consiste em pequenas publicações de até 140 caracteres, conhecidas como *tweet*, que tem como objetivo possibilitar que o usuário se expresse de forma rápida e resumida. No corpo de um *tweet*, o usuário pode fazer uso de marcadores conhecidos como *hashtags* [3], para vincular aquela mensagem à um tópico específico. Com o uso massivo de marcadores em palavras ou frases gera uma grande massa de *tweets* foi criado os *trending topics*, em tradução livre seria os "assuntos mais comentados" onde é mostrado o qual relevante aquele marcador está em determinado lugar, de escolha do usuário, como: Brasil, Mundo, Rio de Janeiro ou outra qualquer localidade. O twitter de acordo com faz uso de *machine learning* para identificar e classificar o idioma da mensagem escrita pelo usuário [4].

2.2 Mineração de opinião

* O que é? * Exemplos no mercado * Etapas (<http://www.inf.ufsc.br/~alvares/INE5644/MineracaoOpin>)

2.3 API

* O que é * APIs mais utilizadas no mundo (case do twitter) * Papel de uma API para integração de serviços (achar referência foda)

2.4 Processamento de linguagem natural

* Linguagem natural (foto da matéria de autômato do Aquino?) * Processamento de linguagem natural * Dificuldades dentro da nossa área de estudo

2.5 Análise de sentimento

* Definição * Objetivo * Premissas * Exemplos e cases de sucesso

2.6 Naive Bayes

* O que é o Naive Bayes * Demonstração matemática do algoritmo * Uso dele em análise de sentimento/classificação

Naive Bayes é um algoritmo probabilístico. Baseado no teorema de Bayes.

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

onde se infere qual é a probabilidade de um evento A dado um evento B. Porém nesse trabalho é utilizado o *Naive Bayes* e sua diferença para o teorema de Bayes é assumir que a posição das palavras que aparecem no texto não importa, daí é acrescentado o *naive*(ingênuo) ao teorema.

Como visto em [5] 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 t_k ocorrer em um documento da mesma classe. O algoritmo tem como objetivo encontrar a melhor classe para um documento maximizando a probabilidade *aposteriori* conforme a equação abaixo, onde n_d é o número de termos no documento d .

$$C_{map} = \operatorname{argmax}_{c \in C} P(c|d) = \operatorname{argmax}_{c \in C} P(c) \prod 1sksn_d P(t_k/d)$$

Capítulo 3

Proposta

Definição da sua proposta.

Se for apresentar os algoritmos use por exemplo:

3.1 Trabalhos relacionados

3.2 Implementação

3.2.1 Crawler

3.2.2 Classificação

3.2.2.1 Algoritmo

3.2.2.2 Construção da base de dados

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 outras bases criadas como: Re-li(referencia), SentiLex-PT(referencia), base da puc(referencia), emoticons(referencia). 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.

3.2.2.3 Massa de treino

3.2.2.4 Massa de teste

3.2.3 Plataforma de análise

Capítulo 4

Resultados e análises

Descreva os resultados encontrados e análises propostas

4.1 Cenários e parâmetros de teste

4.2 Experimentos realizados e resultados

Capítulo 5

Conclusão

Um parágrafo lembrando 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] M. Waite, *Paperback Oxford English dictionary*. Oxford University Press, 2012.
- [4] A. Romann-Kurrik". ("2013", "Fevereiro") "introducing new metadata for tweets". [Online]. Available: "<https://blog.twitter.com/2013/introducing-new-metadata-for-tweets>"
- [5] G. Lucca, I. A. Pereira, A. Prisco, and E. N. Borges, “Uma implementação do algoritmo naïve bayes para classificação de texto,” 2013.