TÍTULO DA TESE

Breno Vieira Arosa

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientadores: Nome do Primeiro Orientador Sobrenome
Nome do Segundo Orientador Sobrenome
Nome do Terceiro Orientador Sobrenome

Rio de Janeiro
Julho de 2019

TÍTULO DA TESE

Breno Vieira Arosa

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Nome do Primeiro Examinador Sobrenome, D.Sc.

Prof. Nome do Segundo Examinador Sobrenome, Ph.D.

Prof. Nome do Terceiro Examinador Sobrenome, D.Sc.

Prof. Nome do Quarto Examinador Sobrenome, Ph.D.

Prof. Nome do Quinto Examinador Sobrenome, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
JULHO DE 2019

Vieira Arosa, Breno

Título da Tese/Breno Vieira Arosa. – Rio de Janeiro: UFRJ/COPPE, 2019.

IX, 13 p.: il.; 29, 7cm.

Orientadores: Nome do Primeiro Orientador Sobrenome
Nome do Segundo Orientador Sobrenome
Nome do Terceiro Orientador Sobrenome

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2019.

Referências Bibliográficas: p. 12 – 12.

1. Primeira palavra-chave. 2. Segunda palavra-chave. 3. Terceira palavra-chave. I. Sobrenome, Nome do Primeiro Orientador *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

A alguém cujo valor é digno
desta dedicatória.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

TÍTULO DA TESE

Breno Vieira Arosa

Julho/2019

Orientadores: Nome do Primeiro Orientador Sobrenome
Nome do Segundo Orientador Sobrenome
Nome do Terceiro Orientador Sobrenome

Programa: Engenharia Elétrica

Apresenta-se, nesta tese, ...

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

THESIS TITLE

Breno Vieira Arosa

July/2019

Advisors: Nome do Primeiro Orientador Sobrenome
Nome do Segundo Orientador Sobrenome
Nome do Terceiro Orientador Sobrenome

Department: Electrical Engineering

In this work, we present ...

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nas últimas duas décadas as redes sociais se tornaram um dos principais meios de comunicação. Esse crescimento, em parte, se justifica pela massificação do acesso a internet incluindo dispositivos móveis como *smartphones* e *tablets*. Também alavancado pelos avanços computacionais e pelo desenvolvimento acelerado de novas técnicas e algoritmos, o aprendizado de máquina, em especial o processamento de linguagem natural, tem essas redes como importante objeto de estudo.

Desde a chamada Revolução Digital observamos um progressivo barateamento e facilitação do uso de dispositivos eletrônicos. À medida que essas tecnologias passaram a ser acessíveis não apenas para as corporações mas também para os indivíduos, houve um crescente processo de digitalização de diversos aspectos de nossas vidas. Com a comunicação não foi diferente. O email, por exemplo, substitui desde os anos 70 operações que até então eram apenas possíveis de forma analógica, como pelo uso de cartas. Nesse contexto, as redes sociais, ou mídias sociais, abordam aspectos diferente da comunicação, uma comunicação mais dinâmica e informal.

Apesar de já existirem desde os anos 90, é com a virada do milênio que as primeiras grandes mídias sociais online aparecem, como *LinkedIn*, *MySpace* e *Orkut*. Desde então há um aumento anual da quantidade de seus usuários. Atualmente, estima-se que 3,5 bilhões de pessoas, ou 45% da população mundial utilize pelo menos uma rede social. Este número torna-se ainda mais interessante quando considerado que 4,4 bilhões de pessoas têm acesso à internet. Portanto, quase 80% dos internautas estão em alguma das mídias sociais. No Brasil esses números se acentuam ainda mais; 70% da população tem acesso à internet e 66% utiliza as redes sociais [1].

Além da alta penetração na sociedade, devido à disponibilidade proporcionada pelos dispositivos móveis, nós consumimos boa parte de nosso tempo nessas redes. No mundo, gasta-se em média 2 horas e 16 minutos por dia. Novamente esse número é ainda superior no Brasil, onde a média é de 3 horas e 34 minutos, sendo o segundo país no mundo a usar por mais tempo as redes, ficando apenas atrás das Filipinas.

Essa forte presença fez com que as mídias sociais não impactassem apenas as co-

municações. Hoje em dia esses meios também são comumente utilizados para busca de relacionamentos, compartilhamento de notícias, divulgação de serviços, atendimento ao público, entre outros. As informações que trafegam nas redes exercem grande influência na formação de opinião das pessoas, seja ela em relação a um produto, a um evento ou até mesmo temas políticos, como pôde-se observar nas eleições pelo mundo nos últimos anos.

Portanto, a análise dessas informações que presentes nas redes é importante para as mais diversos aplicações. Contudo, essa grande quantidade de usuários também se reflete no número de dados provindos das mídias sociais. Dentre as estatísticas de uso do ano de 2018 fornecidas pelas próprias redes sociais, tem-se que, diariamente, 300 milhões de fotos são publicadas no *Facebook*, 5 bilhões de vídeos são vistos no *YouTube*, 43 bilhões de mensagens são enviadas no *WhatsApp* e 100 milhões de usuários interagem pelo *Twitter*.

O massivo volume de dados inviabiliza que essas análises sejam feitas manualmente, se tornando necessário o desenvolvimento de ferramentas capazes de automatizar esse processo. Entram aí as técnicas desenvolvidas pelo campo do Processamento de Linguagem Natural, do inglês *Natural Language Processing* (NLP). Foi a partir do anos 50 que esse termo passou a aparecer como um ramo da Inteligência Artificial. Devido à sua complexidade, a linguagem natural acabou inclusive se tornando um critério de inteligência, como proposto no teste de Turing [2].

O Processamento de Linguagem Natural é uma ampla área de pesquisa, abrangendo diferentes estágios da língua, desde os níveis de abstração mais baixos, como o estudo da fonologia e da sintaxe, quanto os maiores, que lidam com a semântica de determinado conteúdo. Neste ramo, busca-se desenvolver métodos capazes de auxiliar e/ou automatizar tarefas como: o reconhecimento de fala, a análise sintática de frases, a extração de entidades, a segmentação por tópicos etc.

Esse conjunto de ferramentas desenvolvidas pelo NLP são essenciais para tratar o volume de textos gerado pelas redes sociais. Com o avanço de técnicas de *Deep Learning*, nos últimos anos pode-se observar um avanço significativo de desempenho, que permitiu até que tarefas de grande dificuldade, como a automatização de traduções e de aplicações de atendimento ao cliente por conversação, sejam possíveis.

Entretanto, apesar do grande potencial dessas técnicas, as mídias sociais apresentam características que diferenciam seu conteúdo dos tipos textuais que tradicionalmente se analisam em NLP, como artigos e textos jornalísticos. Isso dificulta o processamento dessa informação. Enquanto outras características, apesar de presentes em outras formas de texto são neste caso acentuadas. Em geral, as redes sociais se apresentam como meio de conversação ágil, logo, as mensagem que circulam por elas costumam ser extremamente curtas, com amplo uso de abreviações. Por seu caráter informal, observa-se uma alta taxa de erros gramaticais e uma grande uti-

lização de *emoticons*. Além disso, o fato de serem meios de comunicação globais também ressalta a presença de estrangeirismos. Ademais a dinamicidade das redes sociais faz com que a evolução de sentido das palavras seja acelerada. Esses elementos trazem a necessidade de adaptação ou desenvolvimento de novas técnicas para se reproduzir o sucesso obtido pelas técnicas de processamento de texto em documentos com escrita mais formal e estruturada.

Porém, o principal fator que distingue as informações de redes de outros meios é a forte interligação entre diferentes tipos de mídia, como textos, imagens, áudios, fotos e vídeos. Além de metadados importantes, como localização, data e horário, uma propriedade importante das mídias sociais são os atributos referentes às redes de usuário. Exemplos desses dados são o número de amizades de um usuários da rede, e o número de re-compartilhamentos de uma mensagem. Logo, apesar da capacidade das ferramentas de NLP, existe um conjunto de informações que essas técnicas desconsideram, abrindo espaço para que as abordagens sejam multimodais, ou seja, que tratem de diversas destas propriedades.

De certa maneira, as interações entre usuários são o cerne das mídias sociais. Logo, técnicas que se dispõem a analisar esse tipo de informação também são de grande relevância. Nesse quesito, o campo das Redes Complexas, ou Ciência de Redes, é responsável por estudar os algoritmos e comportamentos observados em grafos que representam sistemas reais, como no caso das redes sociais. Assim como o aprendizado de máquina, essa esfera do conhecimento também apresenta um grande crescimento nos últimos anos, fornecendo um novo leque de tecnologias, de forma a descobrir-se aplicabilidades até então inexploradas. Dentre suas aplicações, que possuem importância para o estudos das mídias sociais, podemos ressaltar, por exemplo, a detecção de comunidades, identificação de principais influenciadores, modelagem de propagação de informação.

Finalmente, estes métodos são meios poderosos de análises de redes sociais, principalmente quando aplicados em complemento à informação textual. A inclusão desses dois últimos elementos na análise se faz necessária porque uma mesma mensagem pode ter conotações diferentes quando escrita por usuários de comunidades com ideias distintas. Por isso, um estudo que também considere esses diferentes tipos de elementos se faz importante.

1.1 Motivação

Dados são considerados um dos bens mais valiosos da atualidade, de forma mesmo a serem chamados de "o novo petróleo". Isso porque, assim como o óleo, os dados são preciosos e precisam ser refinados para terem utilidade. Um dos aspectos dessa transformação pode ser observado na mudança cultural de organizações e empresas

que passam a tomar decisões baseadas em dados e métricas coletadas.

A busca por informação de qualidade sobre um serviço ou produto sempre foi importante para consumidores. Quando não havia as tecnologias que usamos hoje essas pesquisa eram feitas majoritariamente no boca-a-boca ou a partir de revistas especializadas. Com a criação da internet e das redes sociais estas passaram também a exercer essa função, com o beneficio de se encontrar opiniões de forma espontânea e em grande quantidade. As mídias sociais se tornaram um dos principais meios de compartilhamento dessa informação. As empresas, por sua vez, tem a oportunidade de utilizar as opiniões que trafegam nas redes para identificar falhas em suas mercadorias, melhoras sua segmentação, planejar novos produtos, entre outras atividades. Com o fácil acesso a coleta desses dados, as ferramentas capazes de extrair o sentimento dessas mensagens tornam-se fundamentais para viabilizar esse procedimento na escala em que ocorrem.

Apesar das dificuldades inerentes a classificação de mensagens de redes sociais técnicas de aprendizado de máquina, sobretudo *Deep Learning*, e de Redes Complexas apresentam êxito em várias tarefas realizadas sobre elas. Entretanto, o sucesso desses modelos, em geral, dependem da quantidade de dados anotados disponíveis para treinamento. Como o processo de anotação é manual esse passa a ser o gargalo da construção de classificadores de sentimento.

Esse empecilho se torna ainda mais notável quando consideramos aplicações que requerem a análise de múltiplas línguas ou que tenham foco em um tema específico, necessitando criação de bases de dados próprias para cada caso de uso. Esses fatos motivam a elaboração de métodos que sejam independentes de bases de treinamento.

1.2 Objetivo

Esse projeto visa desenvolver um método capaz de formar classificadores de análise de sentimento sem a necessidade de bases de dados de treinamento. Essas análises serão feitas sobre dados de mídia sociais e serão explorados atributos tanto textuais quando de redes de usuários. A principal meta desse trabalho é viabilizar o emprego de modelos complexos e que apresentam melhores desempenhos, sem o custo proveniente da anotação de dados.

Para análise das mensagens será avaliado o impacto da utilização classificadores de *Deep Learning* em comparação a métodos lineares tradicionalmente aplicados em NLP. Diferentes arquiteturas de redes de aprendizado profundo serão experimentadas, como redes convolucionais e redes recorrentes. Além disso, as estratégias de representação de palavras também serão variadas. O processo será feito de maneira semi-supervisionada com supervisão distante para anotação automática dos dados.

Técnicas de Redes Complexas serão aplicadas para caracterização de autores

das mensagens. Modelos como Node2Vec e redes convolucionais de difusão, ambos também baseado em aprendizado de máquina, serão comparados. Neste caso, além de avaliar os modelos entre si, será analisado se adicionar informação do usuários quando aplicada em conjunto com o classificador textual decorre em alguma alteração de performance do sistema.

Concluindo, há um amplo conjunto de estudo aplicando de processamento de linguagem natural em redes sociais. Apesar de menor, a ciência de redes também têm um grande reportório de pesquisa sobre esse meio de comunicação. Este trabalho visa preencher a lacuna de sistemas que não necessitam de investimento em anotação de dados e que abrangem a multimodalidade da informação.

1.3 Organização do Texto

Esse documento é organizado da seguinte maneira:

- O capítulo 2 apresenta o problema da análise de sentimento aplicada em mídias sociais e seus desafios. Esse capítulo contém uma breve revisão bibliográfica de classificadores de análise de sentimento.

- No capítulo 3 as técnicas de processamento de linguagem natural são apresentadas. O capítulo descreve tanto os métodos lineares tradicionalmente aplicados a textos quanto os de *Deep Learning*. São caracterizadas também as diferentes formas de representação numérica de palavras.

- As ferramentas de ciência de redes são apresentadas no capítulo 4. Nesse capítulo são mostrados as diferentes técnicas aplicadas a modelagem de usuários de mídias sociais.

- O capítulo 5 descreve o método proposto para desenvolvimento dos classificadores. São apresentados as etapas de formação de bases de dados, de anotação automática da mesma e de classificação.

- Os resultados dos obtidos pelos experimentos propostos são mostrados no capítulo 6.

- Por fim, o capítulo 7 avalia os resultados obtidos, apresenta as conclusões e enumera possíveis desdobramentos do trabalho realizado.

Capítulo 2

Análise de Sentimento de Redes Sociais

Capítulo 3

Classificadores por Processamento de Linguagem Natural

Capítulo 4

Modelos de Redes Complexas

Capítulo 5

Método

Capítulo 6

Resultados e Discussões

Capítulo 7

Conclusões

Referências Bibliográficas

[1] SOCIAL, W. A. "Digital in 2019: Global Overview". `https://wearesocial.com/global-digital-report-2019`. acessado em 3 de Junho de 2019.

[2] TURING, A. M. "Computing machinery and intelligence", *Mind*, v. 59, n. 236, pp. 433–460, 1950.

Apêndice A

Algumas Demonstrações