Proposta de TCC

Análise de Discurso Político em Comentários da Internet

Aluno: Lucas Helfstein Rocha dos Santos
Curso: Bacharelado em Ciência da Computação

Orientadora: Kelly Rosa Braghetto

Departamento de Ciência da Computação (DCC)
Instituto de Matemática e Estatística (IME)
Universidade de São Paulo (USP)

Abril, 2018

1 Introdução

Com o advento da internet, as pessoas passaram a ter acesso a informação de uma maneira muito acelerada. A quantidade de informações novas a que se tem acesso hoje em dia é uma coisa nunca antes vivenciada, isto fez com que a maneira que as pessoas lidam com a informação seja diferente, buscando informações mais sucintas e de fácil entendimento.

A parcela da população brasileira com acesso a internet vem aumentando nos últimos anos. Cerca de 54% da população brasileira já possui acesso a internet. De 2014 para 2016, o aumento do uso de internet por dispotivos móveis aumentou de 7% para 14%. A maior parte dos usuários utilizam a internet para o envio de mensagens instantâneas (%89) e uso de redes sociais (%78) [1].

Temos, portanto, um grande número de usuários brasileiros que estão começando a usar a internet agora e se acostumando com essa nova dinâmica de acesso à informação. Um grande problema que emerge neste cenário é a disseminação de notícias falsas e a propagação de discursos extremistas e de ódio em comentários [3] [5].

2 Objetivos

Em 2018 teremos eleições no Brasil, será um ano de grande engajamento da população em relação ao tema. Como consequência disto, muitas pessoas estarão se manifestando nas redes sociais e na internet, produzindo um grande volume de dados relacionado à situação política nacional.

Este trabalho tem como objetivo coletar e analisar comentários de notícias com temas relacionados à política, buscando encontrar discursos extremistas e de ódio. Através de técnicas de processamento de dados textuais e de aprendizado de máquina, será criado um classificador para lidar com os os comentários das notícias sobre política mais relevantes nas redes sociais *Twitter* e *Facebook*.

3 Metodologia

A seguir, é descrita a abordagem que será utilizada no desenvolvimento deste trabalho.

3.1 Estudo Bibliográfico sobre Mineração de Opiniões e Levantamento de Ferramentas para o Processamento dos Dados

Nesta etapa do projeto, serão estudadas os métodos e técnicas da computação mais relevantes para a realização do projeto.

Alguns tópicos a serem estudados são:

- Processamento de linguagem natural
- Aprendizado de máquina
- Classificadores
- Mineração de opiniões

Trabalhos como Santos et al. [6], Sun et al. [7] e Mostafa et al. [4] serão utilizados como referência para o trabalho.

As opções de software disponíveis para a aquisição, pré-processamento e análise dos dados serão estudadas com maior profundidade, com foco em plataformas livres e bem estabelecidas na comunidade científica.

Exemplos de ferramentas que serão estudadas são:

- Para aquisição dos dados: Tweepy, Facebook Graph API
- Para pré-processamento dos dados: Jupyter, Julia
- Para mineração dos dados: R, Weka, Python, Scikit-learn

3.2 Monitoramento e aquisição de dados de comentários em notícias com temas políticos

A etapa seguinte do projeto se refere ao monitoramento dos comentários mais relevantes de usuários em notícias muito compartilhadas via Facebook e Twitter para criar um corpus interessante de discurso dos usuários. Aqui, entende-se por relevância a chance de um comentário ser lido pelos demais usuários. Ou seja, se leva em conta a quantidade de reações e comentários para o Facebook, *retweets* e *likes* no Twitter.

Para se obter as notícias mais relevantes, trabalharemos em conjunto com o Grupo de Pesquisa em Políticas Públicas para o Acesso à Informação (GPOPAI) obtendo informações do projeto Monitor do Debate Político no Meio Digital [2]. Este projeto acompanha os principais veículos da imprensa brasileira no *Facebook* mantendo uma base semanal das notícias publicadas e suas respectivas quantidades de compartilhamentos.

Com as informações sobre as páginas mais relevantes, os comentários mais relevantes de cada postagem serão obtidos através da API do *Facebook*. Para o caso do *Twitter*, serão buscados *tweets* que contenham a URL da matéria. Os dados serão armazenados num banco de dados que seja eficiente para os acessos posteriores.

Nesta parte, espera-se encontrar certa dificuldade na coleta dos dados, pois cada rede social possui suas próprias políticas de acesso a dados. Além disso, devido à descoberta recente de problemas de uso indevido de dados de usuários do Facebook, mudanças nas políticas de acesso a esse tipo de dado estão sendo amplamente discutidas e implementadas, a fim de proteger a privacidade dos usuários das redes sociais.

Após efetuar o carregamento e pré-processamento dos dados, será iniciada a etapa de mineração de dados propriamente dita, que consiste em agregar e organizar os dados, procurando por padrões, associações e anomalias relevantes [8].

Nessa etapa, primeiro será feita uma exploração descritiva dos dados utilizando as ferramentas mais básicas de análise estudadas na anteriormente e, em seguida, os dados serão analisados utilizando abordagens mais específicas da área de mineração de opiniões.

3.3 Treinamento do classificador

Nesta fase do projeto os dados serão rotulados para serem utilizados no classificador. Os comentários serão rotulados manualmente de acordo com uma avaliação sobre o discurso empregado, os rótulos serão sobre a orientação política do discurso, o grau de extremismo no comentário e se o discurso é de ódio ou não. Esses rótulos poderão ser alterados até o fim do trabalho caso seja possível obter um melhor desempenho do classificador.

3.4 Aplicação do classificador

Com o classificador treinado pelo conjunto de dados rotulado, o mesmo será testado em conjuntos de dados diferentes para se comparar o desempenho. É esperado que o classificador consiga separar com sucesso discursos que sejam muito extremistas ou que incitem claramente o ódio contra alguma parcela da sociedade.

3.5 Análise e apresentação dos resultados

O desempenho das ferramentas empregadas nas análises será comparado, afim de identificar as mais eficientes para os tipos de dados e o domínio de aplicação considerados no estudo. Interfaces gráficas serão utilizadas para apresentar visualizações amigáveis tanto dos resultados das minerações dos dados quanto do desempenho das ferramentas. Finalmente, será confeccionado um relatório com os resultados completos da pesquisa.

4 Cronograma

As atividades previstas para o trabalho são as enumeradas a seguir; a Tabela 4.1 mostra o cronograma para essas atividades:

1. Estudo das técnicas e ferramentas para mineração de opiniões.
2. Monitoramento e aquisição de dados
3. Treinamento do classificador
4. Aplicação do classificador
5. Análise e apresentação dos resultados

Atividade / Mês	3–4	5–6	7– 8	9–10
Atividade 1	X	X		
Atividade 2	X	X	X	
Atividade 3		X		
Atividade 4		X	X	
Atividade 5			X	X

Tabela 4.1: Cronograma para as atividades previstas para o trabalho.

Referências

[1] *Dobra participação do acesso à internet por dispositivos móveis no Brasil*. http://agenciabrasil.ebc.com.br/geral/noticia/2017-09/dobra-participacao-do-acesso-internet-por-dispositivos-moveis-no-brasil. [Online; accessed 20-April-2018].

[2] *Monitor do debate político no meio digital*. https://www.monitordigital.org/. [Online; accessed 20-April-2018].

[3] *Na web, 12 milhões difundem fake news políticas*. http://politica.estadao.com.br/noticias/geral,na-web-12-milhoes-difundem-fake-news-politicas,70002004235. [Online; accessed 20-April-2018].

[4] Abu-Mostafa, Y. S., M. Magdon-Ismail e H. T. Lin: *Learning From Data*. Learning from Data, 21(4):479–481, 2010, ISSN 1044-3983.

[5] Gnipper, P.: *Uma análise sobre a propagação do ódio pela internet e suas consequências*. https://canaltech.com.br/comportamento/uma-analise-sobre-a-propagacao-do-odio-pela-internet-e-suas-consequencias-100018/. [Online; accessed 20-April-2018].

[6] Santos, R. E. S. *et al.*: *Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: resultados preliminares de um mapeamento sistemático*. Revista de Sistemas e Computação, 4:116–125, 2014.

[7] Sun, S., C. Luo e J. Chen: *A review of natural language processing techniques for opinion mining systems*, vol. 36. Elsevier B.V., 2017, ISBN 8621543451. http://dx.doi.org/10.1016/j.inffus.2016.10.004.

[8] Witten, I. H. e E. Frank: *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.