Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Spatio-Temporal Trend Analysis of the Brazilian Elections based on Twitter Data

Bruno Justino Garcia Praciano

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Orientador
Prof. Dr. -Ing João Paulo Lustosa da Costa

Brasília
2014

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Spatio-Temporal Trend Analysis of the Brazilian Elections based on Twitter Data

Bruno Justino Garcia Praciano

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia da Computação

Prof. Dr. -Ing João Paulo Lustosa da Costa (Orientador)
ENE/UnB

Prof. Dr. Rafael Timóteo de Sousa Júnior
ENE/UnB

Dr. -Ing Ricardo Kerhle Miranda
ENM/UnB

Prof. Dr. José Edil Guimarães de Medeiros
Coordenador do Curso de Engenharia da Computação

Brasília, 24 de dezembro de 2014

Dedicatória

Na *dedicatória* o autor presta homenagem a alguma pessoa (ou grupo de pessoas) que têm significado especial na vida pessoal ou profissional. Por exemplo (e citando o poeta): *Eu dedico essa música a primeira garota que tá sentada ali na fila. Brigado!*

Agradecimentos

Nos *agradecimentos*, o autor se dirige a pessoas ou instituições que contribuíram para elaboração do trabalho apresentado. Por exemplo: *Agradeço aos gigantes cujos ombros me permitiram enxergar mais longe. E a Google e Wikipédia.*

Resumo

O *resumo* é um texto inaugural para quem quer conhecer o trabalho, deve conter uma breve descrição de todo o trabalho (apenas um parágrafo). Portanto, só deve ser escrito após o texto estar pronto. Não é uma coletânea de frases recortadas do trabalho, mas uma apresentação concisa dos pontos relevantes, de modo que o leitor tenha uma ideia completa do que lhe espera. Uma sugestão é que seja composto por quatro pontos: 1) o que está sendo proposto, 2) qual o mérito da proposta, 3) como a proposta foi avaliada/validada, 4) quais as possibilidades para trabalhos futuros. É seguido de (geralmente) três palavras-chave que devem indicar claramente a que se refere o seu trabalho. Por exemplo: *Este trabalho apresenta informações úteis a produção de trabalhos científicos para descrever e exemplificar como utililzar a classe LaTeX do Departamento de Ciência da Computação da Universidade de Brasília para gerar documentos. A classe `UnB-CIC` define um padrão de formato para textos do CIC, facilitando a geração de textos e permitindo que os autores foquem apenas no conteúdo. O formato foi aprovado pelos professores do Departamento e utilizado para gerar este documento. Melhorias futuras incluem manutenção contínua da classe e aprimoramento do texto explicativo.*

Palavras-chave: Big Data, Aprendizado de Máquina Supervisionado, Análise de Sentimentos, Máquina de Vetor de Suporte

Abstract

O *abstract* é o resumo feito na língua Inglesa. Embora o conteúdo apresentado deva ser o mesmo, este texto não deve ser a tradução literal de cada palavra ou frase do resumo, muito menos feito em um tradutor automático. É uma língua diferente e o texto deveria ser escrito de acordo com suas nuances (aproveite para ler `http://dx.doi.org/10.6061%2Fclinics%2F2014(03)01`). Por exemplo: *This work presents useful information on how to create a scientific text to describe and provide examples of how to use the Computer Science Department's LaTeX class. The `UnB-CIC` class defines a standard format for texts, simplifying the process of generating CIC documents and enabling authors to focus only on content. The standard was approved by the Department's professors and used to create this document. Future work includes continued support for the class and improvements on the explanatory text.*

Keywords: Big Data, Supervised Machine Learning, Sentiment Analysis, Support Vector Machine

Contents

List of Figures

Acronyms

AM Aprendizado de Máquina.

API Aplication Programming Interface.

IBGE Instituto Brasileiro de Geografia e Estatística.

TM Text Mining.

Chapter 1

Introduction

Com a popularização da internet tem revolucionado as sociedades com o passar do tempos, pois agora é possível conectar várias pessoas, e realizar trocar de informações em tempo real e com o custo muito baixo em relação aos veículos tradicionais de mídia [2].

As notícias tem sido compartilhadas de maneira muito rápida e eficente e com a utilização massiva das redes de relacionamento, as pessoas podem trocar ideias e opiniões acerca de determinado assunto e com isso facilitar o acesso de todos. Com o passar dos anos as redes sociais já fazem parte da vida de várias pessoas, e com isso as relações interpessoais modificaram-se e esse mundo tem gerado muitos dados de fácil e livre acesso [3].

Com as redes sociais é possível comunicar-se com pessoas de diversas nacionalidades e características, com a grande amplitude que essas alcançam, o volume de dados é algo imensurável e também é uma fonte de dados inesgotável. Com todo esse volume de informações, o ambiente torna-se atrativo para aplicar técnicas de aprendizado de máquina e outros tipos de análises.

1.1 Motivation

De acordo com o IBGE, no Brasil mais de 116 milhões de pessoas tem acesso a internet, ou seja, grande parte da população está expressando suas ideias de forma livre nas redes sociais. E como as eleições em um evento muito importante em qualquer democracia, realizar análise de sentimentos nos textos provenientes de redes sociais se tornam cada vez mais atrativos. O Brasil ocupa a 4ª no ranking de países com a quantidade de pessoas com acesso a internet[4].

Na Figura 1.1 é possível visualizar que grande parte do conteúdo disponível na internet está em inglês, ou seja, é por esse motivo que existem dicionários léxicos para atividades

de TM, como é o caso do WordNet [5], uma das maiores bases de dados do mundo para essa atividade.

As pesquisas utilizando TM com o idioma português ainda são recentes e poucos exploradas, pois a maioria das ferramentas são desenvolvidas para atender o idioma inglês. Mas a análise de sentimento em comentários de redes sociais, pode ser útil em diversas áreas como venda de um produto, avaliação de um estabelecimento, marketing e até mesmo realizar predições de eleições que é o objetivo desse trabalho.

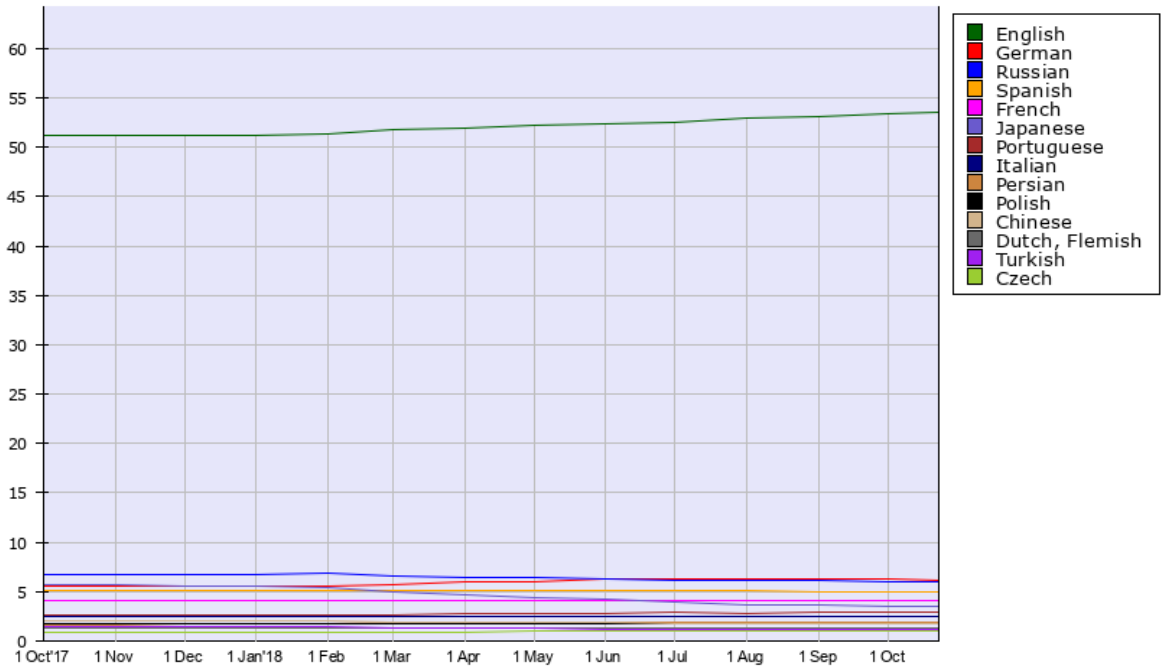

Figure 1.1: Usage of content languages for websites [1].

1.2 Problems

Existem várias redes sociais em funcionamento, e cada uma tem o foco diferente, por exemplo, o *Twitter* é uma rede social com o intuito de formação de opinião, pois grande parte dos usuários a utilizam para compartilhar texto pequenos de até 140 caracteres e também apresenta uma API aberta, mas essa possui limitação em relação ao número de requisições e também ao período que pode efetuar uma busca, que atualmente são de 14 dias [6].

A rede social mais utilizado no Brasil é o *Facebook*, que tem mais de 120 milhões de usuários ativos no Brasil[7], mas após os escândalos das eleições americanas que a envolveram [8] foram adotadas inúmeras medidas de segurança para evitar a extração de

dados da plataforma, atualmente para utilizar a API da empresa é necessário ser aprovado e também conta com o número limitado de requisições que podem ser feitas por cada token de segurança.

É importante citar que os dicionários em língua portuguesa para realizar esse tipo de atividades ainda são pouco desenvolvidos, para esse trabalho foi utilizado dois dicionários em conjunto para melhorar os resultados. O OpLexicon [9] foi combinado com o Sentilex [10], pois atualmente são os melhores dicionários abertos em português para realizar análise de sentimentos, também foi utilizada uma biblioteca chamada TextBlob [11] que é necessário realizar a tradução para o inglês, o que acaba ocasionando um viés na etapa de processamento dos dados.

1.3 Objectives

O objetivo desse trabalho é criar uma forma de predição e um ambiente que expresse a opinião dos usuários da rede social *Twitter* aplicando em textos curtos técnicas de AM, o intuito principal desse trabalho é a utilização do framework proposto em textos que falam sobre políticos que estão concorrendo a cargos eletivos, mas com a inserção de outros textos na fase de treinamento do modelo de AM é possível ampliar as opções e realizar diversas análises para distintas áreas.

1.4 Related work

Em [**?**] os autores definiram o conceito de análise de sentimento em vários níveis e também utilizaram algoritmos que realizam o reconhecimento de entidades, que é a detecção de nomes próprios e com isso remover esses substantivos da análise de sentimento para que o os resultados sejam melhores. O dicionário léxico utilizado para a classificação dos textos provenientes do *Twitter* foi o SentiWordNet, e as polaridades utilizadas nesse trabalho foram três: Positivo, Negativo e Neutro. Foi abordado duas paradigmas de AM, o supervisionado e o não-supervisionado e com o isso o melhor resultado foi de 90% na utilização de algoritmos supervisionados.

No artigo [**?**] foi utilizado um filtro baseado na mineração de opiniões, que é uma das áreas de análise de sentimentos. Foi utilizadas técnicas de decomposição tensorial para capturar interações intrínsecas, pois como o dado é multidimensional, o autor dividiu entre usuários, filmes e outros aspectos, com a aplicação dessas técnicas, houve uma grande redução no esforço computacional do computador na parte de análise de sentimentos, pois o *dataset* utilizado foi reduzido após a decomposição tensorial.

Em [2] os autores aplicaram TM nos dados provenientes do *Twitter* que citavam as eleições presidencias de 2012 da Coréia do Sul. Foram utilizadas distintas técnicas: *topic modeling* para acompanhar as mudanças nos assuntos mais falados do rede sociais, técnicas de análise de rede foram utilizadas para verificar quais pessoas eram citadas e por quem. Os resultados sugeriram que o *Twitter* pode ser um aliado para detectar as mudanças no contexto social enquanto são analisados o texto de quem escreveu.

Em [12] é proposto um sistema para acompanhar as eleições francesas através de tópicos escritos no *Twitter* através da análise de sentimentos. Os resultados obtidos convergiram com os resultados divulgados pelas autoridades da França e foram associados as mudanças de popularidade dos candidatos após a eleição.

Em [13], o dataset utilizado nesse trabalham foi o da eleição colombiana de 2014, técnicas de aprendizado supervisionado foram implementadas e também foram rotuladas previamente usuários que seriam spam. Foi implementado um sistema com o objetivo de investigar o potencial que uma rede social tem de interferir em uma votação, e de acordo com os resultados obtidos, foi possível afirmar que os dados utilizados não foram consistentes.

Em [14] foi usado um dicionário léxico e apenas o algoritmo Naïve Bayes para calcular o sentimento de *tweets* que foram coletados 100 dias antes da eleição americanas de 2016. Os autores classificaram manualmente os textos extraídos do *Twitter*. Os resultados obtids sugerem que essa rede social pode ser considerada ao realizar trabalhos com esse intuito.

1.5 Chapters description

O presente trabalho é apresentado com a seguinte estrutura:

- Capítulo 2: Conceitos em Machine Learning e Mineração de Texto. Apresenta o conjunto de técnicas e metologias que foram necessárias para o desenvolvimento desse trabalho.
- Capítulo 3: Framework proposto. Discorre sobre a metodologia empregada no trabalho e ilustra todos os passos seguidos e necessários para entendidmento do modelo.
- Capítulo 4: Resultados. Nesse capítulo são apresentados os resultados obtidos com a utilização do modelo de aprendizado de máquina proposto e também uma breve justificativa a escolha das ferramentas.
- Capítulo 5: Conclusão. As conclusões sobre o tema são expostas.
- Capítulo 6: Trabalhos Futuros. Na última seção são discutidos quais temas que foram levantados que podem ser aprofundados.

Chapter 2

Concepts on Machine Learning and Text Mining

2.1 Basic Concepts

2.1.1 Training and testing phases

2.1.2 Learning of paradigms

2.1.3 Performance measures

2.2 Machine Learning

2.2.1 Support Vector Machine

2.2.2 Naive Bayes

2.2.3 Decision Trees

2.2.4 Logistic Regression

2.3 Natural Language Processing

2.3.1 Pre-Processing

2.3.2 Tokenization

2.3.3 Stemming

2.3.4 Stop Words

Chapter 3

Proposed Framework

3.1 Crawling and Tweet Extraction

3.2 Data Pre-Processing

3.3 Lexical Dictionary

3.4 Sentiment Classification

3.5 Data Visualization

Chapter 4

Results

4.1 Perfomance evaluation of trend analysis

4.2 N-Fold Cross Validation

4.3 Error evaluation of the sentiment analysis via SVM

4.4 Spatio Trend Analysis

4.5 Election Results

Chapter 5

Conclusion

Este documento serve de exemplo da utilização da classe `UnB-CIC` para escrever um texto cujo objetivo é apresentar os resultados de um trabalho científico. A sequência de ideias apresentada deve fluir claramente, de modo que o leitor consiga compreender os principais conceitos e resultados apresentados, bem como encontrar informações sobre conceitos secundários.

Referências

[1] W3Techs: *Historical trends in the usage of content languages for websites.* https://w3techs.com/technologies/history_overview/content_language. ix, 2

[2] Song, M., M. C. Kim e Y. K. Jeong: *Analyzing the political landscape of 2012 Korean presidential election in Twitter*. IEEE Intelligent Systems, 29(2):18–26, 2014. 1, 4

[3] Araniti, G., I. Bisio e M. De Sanctis: *Towards the reliable and efficient interplanetary internet: A survey of possible advanced networking and communications solutions*. Em *2009 First International Conference on Advances in Satellite and Space Communications*, páginas 30–34, July 2009. 1

[4] Stats, Internet Live: *Elaboration of data by international telecommunication union (itu), united nations population division, internet mobile association of india (iamai), world bank.* http://www.internetlivestats.com/internet-users-by-country/. 1

[5] Miller, G. A.: *Wordnet: a lexical database for English*. Communications of the ACM, 38(11):39–41, 1995. 2

[6] Twitter: *Twitter developer platform*. https://developer.twitter.com. 2

[7] Comunicação, Empresa Brasil de: *Facebook chega a 127 milhões de usuários no brasil*. http://agenciabrasil.ebc.com.br/economia/noticia/2018-07/facebook-chega-127-milhoes-de-usuarios-no-brasil. 2

[8] País, El: *Cambridge analytica, empresa pivô no escândalo do facebook, é fechada.* https://brasil.elpais.com/brasil/2018/05/02/internacional/1525285885_691249.html. 2

[9] Souza, Marlo e Renata Vieira: *Sentiment analysis on twitter data for portuguese language*. Em *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*, PROPOR'12, páginas 241–247, Berlin, Heidelberg, 2012. Springer-Verlag, ISBN 978-3-642-28884-5. http://dx.doi.org/10.1007/978-3-642-28885-2_28. 3

[10] Neuenschwander, Bruna, Adriano C.M. Pereira, Wagner Meira, Jr. e Denilson Barbosa: *Sentiment analysis for streams of web data: A case study of brazilian financial markets*. Em *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, páginas 167–170, New York, NY, USA, 2014. ACM, ISBN 978-1-4503-3230-9. http://doi.acm.org/10.1145/2664551.2664579. 3

[11] Loria, S.: *TextBlob: Simplified text processing.* `http://textblob.readthedocs.io/en/dev/index.html`. 3

[12] Wegrzyn-Wolska, K. e L. Bougueroua: *Tweets mining for French presidential election.* Em *2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN)*, páginas 138–143, Nov 2012. 4

[13] Cerón-Guzmán, J. A. e E. León-Guzmán: *A sentiment analysis system of Spanish tweets and its application in Colombia 2014 presidential election.* Em *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, páginas 250–257, Oct 2016. 4

[14] Joyce, B. e J. Deng: *Sentiment analysis of tweets for the 2016 US presidential election.* Em *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, páginas 1–4, Nov 2017. 4

Appendix A

A.1 Appendix