Universidade Federal
do Rio de Janeiro
Escola Politécnica

ANÁLISE DE NOTÍCIAS DO MERCADO FINANCEIRO UTILIZANDO PROCESSAMENTO DE LINGUAGEM NATURAL E APRENDIZADO DE MÁQUINA PARA DECISÕES DE SWING TRADE

Lucas Gama Canto

Projeto de Graduação apresentado ao Curso de Engenharia de Controle e Automação da Escola Politécnica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Engenheiro.

Orientador: Heraldo Luís Silveira de Almeida

Rio de Janeiro
Março de 2020

ANÁLISE DE NOTÍCIAS DO MERCADO FINANCEIRO UTILIZANDO PROCESSAMENTO DE LINGUAGEM NATURAL E APRENDIZADO DE MÁQUINA PARA DECISÕES DE SWING TRADE

Lucas Gama Canto

PROJETO DE GRADUAÇÃO SUBMETIDO AO CORPO DOCENTE DO CURSO DE ENGENHARIA DE CONTROLE E AUTOMAÇÃO DA ESCOLA POLITÉCNICA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE ENGENHEIRO DE AUTOMAÇÃO.

Examinado por:

Prof. [TODO]Nome do Primeiro Examinador Sobrenome, D.Sc.

Prof. [TODO]Nome do Segundo Examinador Sobrenome, Ph.D.

Prof. [TODO]Nome do Terceiro Examinador Sobrenome, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2020

Gama Canto, Lucas

Análise de Notícias do Mercado Financeiro Utilizando Processamento de Linguagem Natural e Aprendizado de Máquina Para Decisões de Swing Trade/Lucas Gama Canto. – Rio de Janeiro: UFRJ/ Escola Politécnica, 2020.

X, 12 p.: il.; 29, 7cm.

Orientador: Heraldo Luís Silveira de Almeida

Projeto de Graduação – UFRJ/ Escola Politécnica/ Curso de Engenharia de Controle e Automação, 2020.

Referências Bibliográficas: p. 10 – 11.

1. Aprendizado de Máquina. 2. Processamento de Linguagem Natural. 3. Mercado Financeiro. I. Silveira de Almeida, Heraldo Luís. II. Universidade Federal do Rio de Janeiro, Escola Politécnica, Curso de Engenharia de Controle e Automação. III. Título.

Ao povo brasileiro, pela total contribuição em minha graduação.

Agradecimentos

Gostaria de agradecer a todas as pessoas e situações que tornaram este momento possível. Em especial, meus pais Benedita e Manoel, pelo suporte e esforço incondicional em apoiar minha decisão de vir estudar engenharia no Rio de Janeiro, aos professores da graduação, que me fizeram evoluir no âmbito acadêmico, profissional e pessoal, em especial ao meu orientador e professor Heraldo, que não mediu esforços para me ajudar neste trabalho, e aos amigos que me apoiaram e participaram do meu processo de graduação.

Resumo do Projeto de Graduação apresentado à Escola Politécnica/ UFRJ como parte dos requisitos necessários para a obtenção do grau de Engenheiro de Automação.

ANÁLISE DE NOTÍCIAS DO MERCADO FINANCEIRO UTILIZANDO PROCESSAMENTO DE LINGUAGEM NATURAL E APRENDIZADO DE MÁQUINA PARA DECISÕES DE SWING TRADE

Lucas Gama Canto

Março/2020

Orientador: Heraldo Luís Silveira de Almeida

Curso: Engenharia de Controle e Automação

Com o objetivo de automatizar análises fundamentalistas de mercado, o uso de tecnologia para processamento de texto vem sendo utilizado constantemente no meio acadêmico[1] e profissional[2]. De forma a contribuir para este campo em crescimento, este trabalho discorre um estudo acerca da criação de modelos preditivos sobre a valorização ou desvalorização de ações na bolsa de valores do Brasil (B3, antiga Bovespa) a partir de notícias sobre o mercado brasileiro de forma a auxiliar decisões de Swing Trade, ou seja, compra e venda de ações dentro de uma janela de tempo maior que um dia.

Para isto, o presente projeto utiliza o framework PyText, que se baseia em conceitos de Aprendizado de Máquina, Redes Neurais e Processamento de Linguagem Natural de forma a desenvolver modelos preditivos com a tarefa de classificação textual.

Abstract of Undergraduate Project presented to POLI/UFRJ as a partial fulfillment of the requirements for the degree of Engineer.

FINANCIAL MARKET NEWS ANALYSIS USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING FOR SWING TRADE DECISIONS

Lucas Gama Canto

March/2020

Advisor: Heraldo Luís Silveira de Almeida

Course: Automation and Control Engineering

In order to automate fundamental market analysis, the use of text processing technology has been constantly used in academic[1] and professional[2] means. To contribute to this growing field, this paper discusses a study about the creation of predictive models regarding the valuation or devaluation of shares on the Brazilian stock exchange (B3, former Bovespa) based on news about the Brazilian market in order to assist Swing Trade decisions, that is, buying and selling stocks within a time window longer than one day.

To this end, the present project uses the PyText framework, which is based on Machine Learning, Neural Networks and Natural Language Processing concepts in order to develop predictive models with the task of textual classification.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Tema

O tema deste trabalho se resume no estudo da criação de modelos preditivos de modo que estes possam prever a valorização ou desvalorização de ações da bolsa de valores por meio do processamento de notícias do mercado brasileiro.

Deste modo, o problema a ser abordado é a identificação de quando uma notícia pode impactar positivamente ou negativamente a variação de preço de ações de forma automatizada.

1.2 Delimitação

Este trabalho se restringe ao processamento de texto em português brasileiro, tendo como foco a predição da variação de preço das ações que fazem parte da bolsa de valores do Brasil, a B3. Pela indisponibilidade de dados sobre notícias brasileiras contendo a informação do horário de lançamento da notícia, o projeto mira em predições dentro de uma janela de tempo maior que um dia, de forma a auxiliar decisões de Swing Trade, isto é, operações de compra e venda de ações numa janela de tempo maior que um dia.

Além disso, o estudo se baseia na ferramenta PyText, um framework recentemente desenvolvido pelo Facebook que providencia modelos de processamento de linguagem natural de última geração através de uma interface simples e extensível[3].

1.3 Justificativa

Diante do crescente número de investidores na bolsa de valores no Brasil, nota-se uma maior preocupação da população brasileira acerca da busca por independência financeira e fontes alternativas de renda com o intuito de contribuir à economia familiar, previdência, ou mesmo utilizar este método como fonte principal de renda[4].

Ao mesmo tempo, estudos associados à inteligência artificial, aprendizado de máquina e processamento de linguagem natural continuam emergindo no meio acadêmico e auxiliando o meio profissional como nunca antes, incluindo o mercado financeiro[5].

Através destes dois fatores, o presente trabalho busca contribuir para a difusão do estudo e uso de algumas destas tecnologias sobre um assunto que gradualmente se encontra dentro do interesse da população brasileira e que colabora para uma possível instauração de uma cultura de economia e independência financeira no Brasil.

1.4 Objetivos

O objetivo geral do presente trabalho é de analisar modelos preditivos associados ao mercado financeiro que possam ser construídos a partir do framework PyText, tendo como objetivos específicos, apresentar: (1) A busca por dados de notícias e do histórico da bolsa de valores; (2) A lógica utilizada para a união destes dados de forma a construir os conjuntos de dados utilizados no treinamento dos modelos; (3) O pré-processamento dos conjuntos de dados; (4) As possíveis configurações do framework utilizado de forma a obter a melhor performance; (5) O detalhamento e a análise dos modelos finais encontrados.

1.5 Metodologia

O trabalho teve início a partir da procura por bases de dados de notícias associadas ao mercado brasileiro e escritas em português do Brasil, seguida pela obtenção do histórico das variações de preço dos ativos que compõem o índice Bovespa. Após isto, o histórico foi filtrado de forma a manter as informações dos 5 índices mais significativos e das varições destes ativos que ocorreram dentro da mesma janela de tempo das notícias obtidas. Em seguida, estes dados foram unidos de forma a obter 5 conjuntos de dados para cada ativo, cada um levando em consideração uma diferente janela de tempo para indicar a valorização: de 1 a 5 dias.

Logo após, houve a etapa de pré-processamento do corpo das notícias de forma a remover possíveis ruídos e facilitar a etapa de treinamento, sem perda de contexto do conteúdo. Com os conjuntos de dados prontos, foram feitos testes no PyText com o objetivo de definir a melhor configuração possível para a natureza dos dados, e assim obter a melhor performance.

Por fim, os testes finais de cada modelo gerado foi detalhado e analisado para permitir uma conclusão e avaliação do processo como um todo.

1.6 Descrição

O capítulo 2 apresenta toda a fundamentação teórica utilizada como base para o projeto a partir de uma breve descrição de como a bolsa de valores funciona e como pode-se obter lucro a partir da mesma, seguida de explicações sobre Aprendizado de Máquina, Processamento de Linguagem Natural, Redes Neurais e o framework Pytext.

No capítulo 3 é detalhado todo o processo executado para obtenção do conjunto de notícias e do histórico da B3, seguido do pré-processamento realizado nestes dois conjuntos e a criação dos conjuntos de dados finais utilizados para o treino, cada um associado a um ativo e uma janela de tempo específica.

Os detalhes das configurações utilizadas no PyText e o treinamento em si é especificado no capítulo 4, onde há uma discussão acerca dos parâmetros encontrados para a geração de modelos mais performáticos, além das métricas finais encontradas para cada modelo gerado.

Por fim, o capítulo 5 apresenta uma conclusão acerca dos modelos encontrados seguido por sugestões que futuramente podem ser aplicadas para a evolução do tema e uma possível melhora de desempenho dos modelos preditivos.

O código desenvolvido para o pré-processamento e geração dos conjuntos de dados e arquivos de configurações do PyText utilizados para a geração dos modelos podem ser encontrados no repositório do github referenciado em [6].

Capítulo 2

Fundamentação Teórica

2.1 Economia Financeira

2.2 Aprendizado de Máquina

2.3 Processamento de Linguagem Natural

2.4 Redes Neurais

2.5 PyText

Capítulo 3

Obtenção e Pré-processamento de Dados

Capítulo 4

Treinamento

Capítulo 5

Conclusões

Segundo a norma de formatação de teses e dissertações do Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia (COPPE), toda abreviatura deve ser definida antes de utilizada.

Do mesmo modo, é imprescindível definir os símbolos, tal como o conjunto dos números reais $\mathbb{R}$ e o conjunto vazio $\emptyset$.

Você deve selecionar seu curso de engenharia usando o comando `\department{Sigla}` e no lugar de Sigla inserir a sigla referente ao seu curso de engenharia. A tabela 5.1 relaciona as siglas dos cursos de engenharia da Escola Politécnica da Universidade Federal do Rio de Janeiro (POLI-UFRJ), enquanto que a tabela 5.2 relaciona as siglas dos programas de pós graduação da COPPE.

Tabela 5.1: Siglas dos cursos de engenharia da Escola Politécnica da UFRJ.

Sigla	Curso
EA	Engenharia Ambiental
ECV	Engenharia Civil
ECI	Engenharia de Computação e Informação
ECA	Engenharia de Controle e Automação
EMAT	Engenharia de Materiais
EPT	Engenharia de Petróleo
EPR	Engenharia de Produção
EEC	Engenharia Eletrônica e de Computação
EET	Engenharia Elétrica
EMC	Engenharia Mecânica
EMET	Engenharia Metalúrgica
ENO	Engenharia Naval e Oceânica
ENU	Engenharia Nuclear

Note também que todas as figuras ou tabelas devem ser citadas no texto. Como ocorre com as tabelas 5.1 e 5.2. Para ilustrar o uso de figuras em LaTeX, considere as figuras 5.1 e 5.2.

Tabela 5.2: Siglas dos programas de pós graduação da COPPE.

Sigla	Curso
PEB	Engenharia Biomédica
PEC	Engenharia Civil
PEE	Engenharia Elétrica
PEM	Engenharia Mecânica
PEMM	Engenharia Metalúrgica e de Materiais
PEN	Engenharia Nuclear
PENO	Engenharia Oceânica
PPE	Planejamento Energético
PEP	Engenharia de Produção
PEQ	Engenharia Química
PESC	Engenharia de Sistemas e Computação
PET	Engenharia de Transportes

Figura 5.1: Logotipo da POLI-UFRJ.

Figura 5.2: Logotipo da COPPE.

Capítulo 6

Revisão Bibliográfica

Para ilustrar a completa adesão ao estilo de citações e listagem de referências bibliográficas, a Tabela 6.1 apresenta citações de alguns dos trabalhos contidos na norma fornecida pela CPGP da COPPE, utilizando o estilo numérico.

Tabela 6.1: Exemplos de citações utilizando o comando padrão `\cite` do LaTeX e o comando `\citet`, fornecido pelo pacote `natbib`.

Tipo da Publicação	`\cite`	`\citet`
Livro	[7]	EXEMPLO-ABRAHAM *et al.* [7]
Artigo	[?]	?]
Relatório	[8]	MAESTRELLO [8]
Relatório	[9]	GARRET [9]
Anais de Congresso	[10]	GURTIN [10]
Séries	[11]	COWIN [11]
Em Livro	[12]	EDWARDS [12]
Dissertação de mestrado	[13]	TUNTOMO [13]
Tese de doutorado	[14]	JUNIOR e R. [14]

É importante notar que, segundo a Norma para a Elaboração Gráfica do Projeto de Graduação da Escola Politécnica da UFRJ para trabalhos de conclusão de curso de engenharia de julho de 2012, as referências bibliográficas podem ser apresentadas de duas formas: (*i*) Referências numeradas e (*ii*) Referências em ordem alfabética. Para exibição numerada, em que a exibição das referências bibliográficas segue a ordem de citação usada no texto, use o comando `\bibliographystyle{coppe-unsrt}`. Para exibição de referências bibliográficas em ordem alfabética, basta usar o comando `\bibliographystyle{coppe-plain}` ao final do documento.

Referências Bibliográficas

[1] LIU, Z., ZHU, H., CHONG, T. Y. "An NLP-PCA Based Trading Strategy On Chinese Stock Market", *Advances in Social Science and Education and Humanities Research*, v. 334, n. 2, pp. 80–89, jul. 2019.

[2] SEDLAK, M. "How Natural Language Processing is transforming the financial industry". `https://www.ibm.com/blogs/watson/2016/06/natural-language-processing-transforming-financial-industry-2/`, 2016. Acessado em Dezembro/2019.

[3] ALY, A., LAKHOTIA, K., ZHAO, S., et al. "PYTEXT: A SEAMLESS PATH FROM NLP RESEARCH TO PRODUCTION", 2018.

[4] DO PAVINI, A. "Cresce número de pessoas físicas como profissionais na Bolsa". `https://exame.abril.com.br/seu-dinheiro/cresce-numero-de-pessoas-fisicas-como-profissionais-na-bolsa/`, 2019. Acessado em Dezembro/2019.

[5] BACHINSKIY, A. "The Growing Impact of AI in Financial Services: Six Examples". `https://towardsdatascience.com/the-growing-impact-of-ai-in-financial-services-six-examples-da386c0301b2`, 2019. Acessado em Dezembro/2019.

[6] CANTO, L. G. "Stock Market Predictor". `https://github.com/lgcanto/stock-market-predictor/`, 2019. Acessado em Dezembro/2019.

[7] EXEMPLO-ABRAHAM, R., MARSDEN, J. E., RATIU, T. *Manifolds and Tensor Analysis and and Applications*. 2 ed. New York, Springer-Verlag, 1988.

[8] MAESTRELLO, L. *Two-Point Correlations of Sound Pressure in the Far Field of a Jet: Experiment*. NASA TM X-72835, 1976.

[9] GARRET, D. A. *The Microscopic Detection of Corrosion in Aluminum Aircraft Structures with Thermal Neutron Beams and Film Imaging Methods*. In: Report NBSIR 78-1434, National Bureau of Standards, Washington and D.C., 1977.

[10] GURTIN, M. E. "On the nonlinear theory of elasticity". In: *Proceedings of the International Symposium on Continuum Mechanics and Partial Differential Equations: Contemporary Developments in Continuum Mechanics and Partial Differential Equations*, pp. 237–253, Rio de Janeiro, ago. 1977.

[11] COWIN, S. C. "Adaptive Anisotropy: An Example in Living Bone". In: *Non-Classical Continuum Mechanics*, v. 122, *London Mathematical Society Lecture Note Series*, Cambridge University Press, pp. 174–186, 1987.

[12] EDWARDS, D. K. "Thermal Radiation Measurements". In: Eckert, E. R. G., Goldstein, R. J. (Eds.), *Measurements in Heat Transfer*, 2 ed., cap. 10, New York and USA, Hemisphere Publishing Corporation, 1976.

[13] TUNTOMO, A. *Transport Phenomena in a Small Particle with Internal Radiant Absorption.* Ph.D. dissertation, University of California at Berkeley, Berkeley and California and USA, 1990.

[14] JUNIOR, P., R., H. *Influência da Espessura da Camada Intrínseca e Energia do Foton na Degradação de Células Solares de Silício Amorfo Hidrogenado.* Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro and RJ and Brasil, 1994.

Apêndice A

Algumas Demonstrações