

Instituto de Pesquisas Tecnológicas do Estado de São Paulo

André Hideo Hayashi

**Processo para predição de preços das ações no mercado
financeiro com uso de *Big Data***

**São Paulo
2017**

André Hideo Hayashi

Processo para predição de preços das ações no mercado financeiro com uso de *Big Data*

Dissertação de Mestrado apresentada ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo - IPT, como parte dos requisitos para a obtenção do título de Mestre em Engenharia de Computação: Redes de Computadores

Data da aprovação ____/____/____

Prof. Dr. Alexandre Barbieri (Orientador)
IPT – Instituto de Pesquisas Tecnológicas
do Estado de São Paulo

Membros da Banca Examinadora:

Prof. Dr. Alexandre José Barbieri de Sousa (Orientador)
IPT – Instituto de Pesquisas Tecnológicas do Estado de São Paulo

Prof. Dr. Roberto Kenji Hiramatsu (Membro)
POLI – Escola Politécnica da USP

Prof. Dr. Mario Menezes (Membro)
IPEN – Instituto de Pesquisas Energéticas e Nucleares

André Hideo Hayashi

Processo para predição de preços das ações no mercado financeiro
com uso de *Big Data*

Dissertação de Mestrado apresentada ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo - IPT, como parte dos requisitos para a obtenção do título de Mestre em Engenharia da Computação.

Área de Concentração: Redes de Computadores

Orientador: Prof. Dr. Alexandre J. Barbieri

São Paulo
Agosto / 2017

Ficha Catalográfica

Elaborada pelo Departamento de Acervo e Informação Tecnológica – DAIT
do Instituto de Pesquisas Tecnológicas do Estado de São Paulo - IPT

H412p

Hayashi, André Hideo

Processo para predição de preços das ações no mercado financeiro com uso de Big Data. / André Hideo Hayashi. São Paulo, 2017.

87p.

Dissertação (Mestrado em Engenharia de Computação) - Instituto de Pesquisas Tecnológicas do Estado de São Paulo. Área de concentração: Redes de Computadores.

Orientador: Prof. Dr. Alexandre J. Barbieri

DEDICATÓRIA

Dedico a minha família que sempre compreendeu a necessidade de sacrifícios para a conclusão deste trabalho.

AGRADECIMENTOS

Ao meu orientador Prof. Dr. Alexandre José Barbieri de Sousa.

Aos membros da banca professores Dr. Mario Menezes e Dr. Roberto Kenji Hiramatsu.

Às empresas e pessoas que, de forma direta ou indireta, contribuíram para o desenvolvimento deste trabalho.

RESUMO

O mercado financeiro é muito inconstante, entretanto, os investidores têm a difícil tarefa de acompanhar e tentar prever as oscilações para que suas estratégias resultem em melhores retornos financeiros. Com a utilização do *Big Data* e estatísticas matemáticas bayesianas o investidor pode acompanhar as notícias financeiras que afetam os principais indicadores financeiros, de forma contínua, e tentar prever as oscilações de preços das ações. Muitos sistemas eletrônicos disponibilizados por bancos e corretoras baseiam-se em informações de liquidez diárias dos ativos das empresas participantes da Bolsa e, principalmente, em gráficos de tendências e poucos sistemas realizam predições com *Naive Bayes*, que utiliza um modelo probabilístico baseado no conhecimento prévio, além de exemplos de treinamento para determinar a probabilidade de uma hipótese. O objetivo deste trabalho é propor um processo de predição de preços das ações de uma empresa listada no Ibovespa, auxiliando um investidor para um melhor momento de compra ou venda de ações. Nesse sistema preditivo será utilizado *Big Data*, *Naive Bayes* e o processamento de linguagem natural. Será observado em um experimento o resultado do processo de predição e comparado com o preço das ações durante o pregão eletrônico.

Palavras Chaves: Palavras chave: *Big Data*, *Naive Bayes*, processamento de linguagem natural, mercado financeiro.

ABSTRACT

Process for prediction of stock prices in the financial market with the use of Big Data

The financial market is very fickle, however, investors have the difficult task of tracking and trying to predict the swings so that their strategies result in better financial returns. With the use of Big Data and Bayesian mathematical statistics the investor can track the financial news that affect the key financial indicators on an ongoing basis, and try to predict stock price swings. Many electronic systems provided by banks and brokerage firms are based on daily liquidity information on the assets of the companies participating in the Exchange and mainly on trend charts and few systems make predictions with Naive Bayes using a probabilistic model based on prior knowledge, As well as examples of training to determine the likelihood of a hypothesis. The objective of this work is to propose a process of prediction of stock prices of a company listed on the Ibovespa, helping an investor to better buy or sell shares. In this predictive system will be used Big Data, Naive Bayes and the Natural Language Processing. It will be observed in an experiment the result of the prediction process and compared with the stock price during the electronic trading session.

Keywords: Big Data, Naive Bayes, Natural Language Processing, financial market

Lista de ilustrações

Fig. 1 - Fases do método de trabalho	16
Fig. 2 - Exemplo do cálculo de probabilidade de um valor das ações da empresa Ambev durante o pregão eletrônico.	19
Fig. 3 - Linhas de tendências de mercado.	26
Fig. 4 - Confirmação da tendência de alta e baixa, após a união do terceiro ponto.....	26
Fig. 5 - Análise de tendência. União do terceiro ponto e o preço das ações não se mantiveram em alta.	27
Fig. 6 - MapReduce processando com o computador Mestre e Escravos.....	32
Fig. 7 - Arquitetura do Big Data HDFS e MapReduce com Mestre e 01 Escravo.	39
Fig. 8 - Etapas do processo de classificação de textos com PLN.....	41
Fig. 9 - Coleta de informações de sites para composição de aferição de predição dos preços da ação da Ambev.....	46
Fig. 10 - Fluxograma da média de volume das ações negociadas.....	48
Fig. 11 - Coleta de preços e volume da ação Abev3 e Ibovespa do site BrInvesting para composição de aferição de predição dos preços da ação da Ambev.....	49
Fig. 12 - Utilização da função transpose da linguagem e ambiente de estatísticas R transpondo a matriz de preço da ação Abev3 para data frame.....	51
Fig. 13 - Gráfico com os valores da ação da Abev3 de 03 a 19/05 de 2016.....	54
Fig. 14 - Exemplo de cálculo da probabilidade do preço da ação Abev3 subir no período analisado 03 a 19/05 de 2016.....	56
Fig. 15 - Cálculo da probabilidade do preço da ação Abev3 subir no período analisado 11.08.2016 a 09.09.2016.....	58
Fig. 16 - Processo macro de predição pela linguagem e ambiente de estatísticas R.....	60
Fig. 17 - Etapas para a realização da predição com a função naiveBayes.....	60
Fig. 18 - Fragmento de código para processamento da predição com a função naiveBayes pela linguagem e ambiente de estatísticas R e pacote E1071.....	62
Fig. 19 - Apresentação da função naiveBayes pela linguagem e ambiente de estatísticas R e pacote E1071.....	63
Fig. 20 - Preço da ação Abev3 em 20/09/2016 das 10 às 17:00.....	64

Fig. 21 - Fluxograma para o cálculo de predição.	66
Fig. 22 - Gráfico comparativo valor real da ação Abev3 e valor da predição.	69
Fig. 23 - Gráfico do dia 14/02/2017 apresentando o ajuste de R\$ 0,02 à predição aproximando-se do valor da ação Abev3.	71
Fig. 24 - Gráfico do dia 10/02/2017 apresentando o ajuste de R\$ 0,03 à predição depois de verificado a média do volume de ações.	72

Lista de tabelas

Tabela 1 - Trabalhos relacionados.....	32
Tabela 2 - Exemplo de classificação de textos atribuídos a baixa e alta probabilidade para composição da predição do índice FTSE100.....	33
Tabela 3 - Comparativo de eficiência individual dos classificadores KNN, SVM e Naive Bayes.	35
Tabela 4 - Comparativo de tempo de processamento de 100 arquivos pequenos e 1 arquivo único.....	44
Tabela 5 - Dados do preço da ação da Ambev transportados para uma planilha para ilustrar a acomodação de preço do dia anterior.....	50
Tabela 6 - Tabela com os valores da ação da Abbev3 de 03 a 19/05 de 2016 coletados às 11h.	53
Tabela 7 - Tabela de frequência para composição da predição do valor da ação da Abbev3 de 03 a 19/05 de 2016.....	54
Tabela 8 - Tabela de probabilidade para composição da predição do valor da ação da Abbev3 de 03 a 19/05 de 2016.....	55
Tabela 9 - Tabela com os valores da ação da Abbev3 de 11.08.2016 a 09.09.2016 às 10h37min.....	57
Tabela 10 - Tabela de probabilidade para composição da predição do valor da ação da Abbev3 de 11.08.2016 a 09.09.2016.....	58
Tabela 11 - Diferença do último e penúltimo valor do 12:05 às 12:02.....	65
Tabela 12 - Análise do histórico dos valores dos preços da ação e da predição.....	68

Lista de fórmulas

Fórmula 1 - Classificador Naive Bayes.	18
Fórmula 2 - Fórmula de aferição de preços.	20
Fórmula 3 - Classificador Naive Bayes.	29
Fórmula 4 - Fórmula de aferição de preços.	47
Fórmula 5 - Classificador Naive Bayes com variáveis C e X.	52
Fórmula 6 - Cálculo de predição.	65

Lista de abreviaturas e siglas

Ambev	Companhia de Bebidas das Américas
Abev3	Código da ação Ambev no BM&FBovespa
BM&FBovespa	Bolsa de Valores, Mercadorias e Futuros de São Paulo
<i>Dividend Yield</i>	Dividendos / preço da ação
EBITDA	<i>Earning Before Interest, Taxes, Depreciation and Amortization</i>
FTSE100	<i>Financial Times Stock Exchange</i>
FTP	<i>File Transfer Protocol</i>
HDFS	<i>Hadoop Distributed File System</i>
HTML	<i>Hipertext Markup Language</i>
HTTP	<i>Hypertext Transfer Protocol</i>
HTTPS	<i>Hyper Text Transfer Protocol Secure</i>
KNN	<i>K-Nearest Neighbor</i>
LPA	Lucro por ação
<i>MapReduce</i>	<i>Map e Reduce</i>
MSN	<i>Microsoft Network</i>
NNTP	<i>Network News Transfer Protocol</i>
NB	<i>Naive Bayes</i>
P/L	Preço/Lucro
P/VPA	Preço da ação / valor patrimonial por ação
POP3	<i>Post Office Protocol Version 3</i>
R	<i>Comprehensive R Archive Network</i>
RAM	Random Access Memory
RPL	Retorno sobre patrimônio líquido
SMTP	<i>Simple Mail Transfer Protocol</i>
SQL	Structure Query Language

SSH	<i>Secure Shell</i>
SVM	<i>Support Vector Machine</i>
S&P500	<i>Standard & Poor's</i>
VE/Ebitda	Valor de empresa sobre Ebitda

SUMÁRIO

1 INTRODUÇÃO.....	14
1.1 Motivação	14
1.2 Objetivo	15
1.3 Contribuições	15
1.4 Método de trabalho	16
1.5 Organização do trabalho	20
2 ESTADO DA ARTE.....	22
2.1 Introdução.....	22
2.2 Mercado financeiro	22
2.3 Análise técnica, análise fundamentalista e previsões	22
2.4 Probabilidades e estatísticas	28
2.5 Teorema de Bayes.....	29
2.6 Big Data	30
2.7 Map Reduce	31
2.8 Trabalhos relacionados.....	32
2.9 Conclusões da seção.....	37
3 DESENVOLVIMENTO	38
3.1 Ambiente Big Data	38
3.2 Arquitetura Big Data.....	38
3.3 Programação com Map e Reduce e Processamento de Linguagem Natural.....	39
3.4 Programa coletor de dados de sites.....	45
3.5 RStudio para cálculo de aferição	46
3.6 Coleta dos preços das ações	48
3.7 Programa coletor dos preços das ações da Ambev e volume das ações negociadas....	49
3.8 Análise de linhas.....	50
3.9 Predição do valor da ação ABEV3 com classificador Naive Bayes.....	51
3.10 Predição com linguagem e ambiente de estatísticas R	59
3.11 Detalhamento da função naiveBayes.....	60
3.12 Cálculo de predição	63
4 AVALIAÇÃO DOS RESULTADOS.....	67
4.1 Resultado da predição Naive Bayes	67
4.2 Resultado do Big Data e processamento de linguagem natural	70
4.3 Resultado	73
5 CONCLUSÃO	74
5.1 Trabalhos Futuros.....	75
6 REFERÊNCIAS	76

7 ANEXOS.....	79
---------------	----

1 INTRODUÇÃO

1.1 Motivação

Apesar de ser controlado por vários órgãos regulamentadores, o mercado financeiro é muito inconstante. Inconstância que pode ser influenciada por vários aspectos como crise econômica, mudanças políticas e governamentais, alterações no mercado exterior, fusões e aquisições de empresas, catástrofes naturais ou não e outros fatores econômicos.

Nesse cenário, os investidores têm a difícil tarefa de acompanhar, de forma contínua, as oscilações do mercado para que suas estratégias resultem em melhores retornos financeiros. Não é uma tarefa simples acompanhar as diversas informações veiculadas em jornais, revistas, televisão, *Internet* ou outros canais, depois filtrá-las e analisá-las para tomar uma decisão.

Alguns investidores realizam as consagradas análises técnica e fundamentalista, utilizando ferramentas gráficas para acompanhar e até mesmo antecipar os riscos do mercado.

Muitos sistemas eletrônicos, disponibilizados por corretoras e instituições bancárias para seus clientes, baseiam-se nas informações de liquidez diárias das empresas participantes da Bolsa e em gráficos de tendências, sendo que, apenas em alguns casos, realizam previsões baseados no modelo *Bayesiano*.

Prever o futuro é o desejo de muitos investidores e não existem garantias de que um sistema consiga fazê-lo de forma permanente e com certezas de acertos, mas é possível prever o mercado com a utilização de probabilidades e estatísticas.

Para os investidores as informações financeiras são instrumentos do trabalho diário. Eles as procuram em diferentes fontes e sites para tentar acompanhar as oscilações do mercado tentando, em alguns casos, antecipar as ações que possam ajudá-los a realizar seus investimentos. É possível, ainda, coletar informações de sites financeiros por meio de programas extratores e processá-los no *Big Data*, gerando estatísticas financeiras.

A probabilidade e a estatística possibilitam a utilização de um modelo matemático capaz de calcular, por exemplo, os valores a que podem chegar os preços das ações de empresas. Essas predições auxiliam pessoas ou empresas a tomarem decisões de compras ou vendas de ações, ou mesmo a preverem indicadores econômicos.

Um modelo muito usado para predições é o modelo *Bayesiano*, fundamentado no teorema de *Bayes*. *Naive Bayes* utiliza um classificador probabilístico que obedece regras próprias para realizar predições. (SANTIDHANYAROJ P, 2015).

É possível também utilizar o *Naive Bayes* para classificar textos oriundos dos principais sites de notícias na Internet. Com isso, é possível estimar uma probabilidade de que determinada notícia poderá influenciar uma predição. (SHYNKEVICH Y, 2015).

O uso de artigos financeiros com o processamento de linguagem natural pode ajudar a identificar os sentimentos positivos e negativos dos textos. (KIM, 2014). Assim, a classificação de textos de sites de notícias pode facilitar o acompanhamento das oscilações dos preços das ações.

1.2 Objetivo

O objetivo deste trabalho é propor um processo de predição de preços das ações no mercado financeiro, com o uso de um sistema preditivo, usando a extração de textos de sites de notícias financeiras, *Big Data* e processamento de linguagem natural para tratamento e análise dos sentimentos presentes nos textos que podem influenciar os preços das ações. Neste processo, será utilizado o classificador *Naive Bayes* para predição dos valores das ações.

Para testar o processo de predição, será desenvolvido um sistema que fará predições e comparará o resultado obtido com as oscilações dos preços das ações durante o pregão eletrônico.

1.3 Contribuições

Este trabalho contribui para a utilização do classificador *Naive Bayes* em predições da bolsa de valores IBovespa, especialmente, no uso do *Big Data*, pouco

explorado na bibliografia levantada. Adicionalmente serão aprofundados estudos sobre:

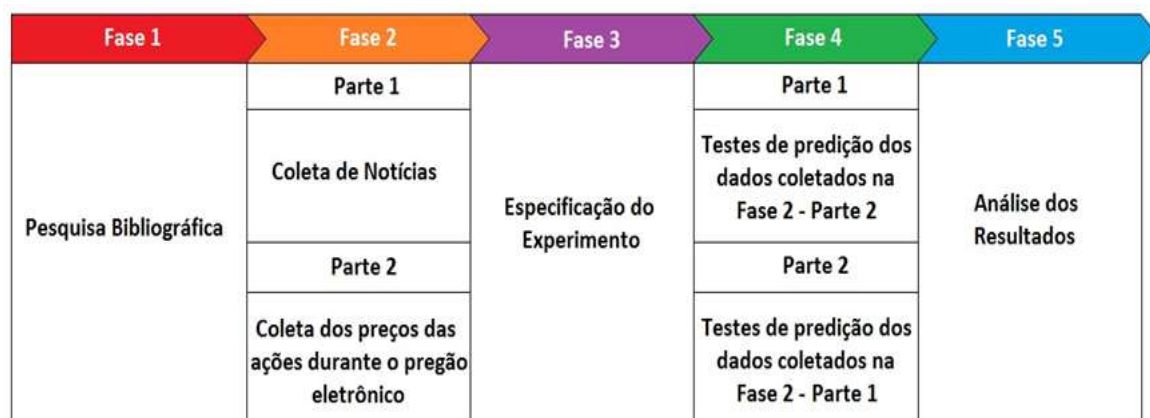
- predição com *Naive Bayes*;
- classificação de textos com a utilização de processamento de linguagem natural;
- processo de predição dos preços das ações com *Naive Bayes* e utilização do *Big Data* como ferramentas de análise financeira ou ainda adaptadas para outro segmento.

1.4 Método de trabalho

Inicialmente, será realizada uma pesquisa exploratória, visando prover o pesquisador de um conhecimento maior sobre o tema ou problema em análise, uma vez que no início da investigação o conhecimento, a compreensão e a familiaridade do pesquisador são insuficientes ou inexistentes. (MATTAR, 1994).

O método de trabalho foi dividido em cinco fases, sendo a fase 1 a pesquisa bibliográfica; a fase 2, o levantamento de dados; a fase 3, a especificação do experimento; a fase 4 a execução dos testes e, a fase 5 as análises dos resultados, conforme a figura 1.

Fig. 1 - Fases do método de trabalho



Fonte: Autor.

Fase 1 – Pesquisa Bibliográfica

Nesta fase foram realizadas pesquisas em livros e artigos relacionados à predição da bolsa de valores, com a utilização do *Naive Bayes*, do *Big Data*, do

processamento de linguagem natural e da linguagem e ambiente de estatísticas R. Essas referências são utilizadas para fundamentação teórica deste trabalho.

Fase 2 – Levantamento de dados

Nessa fase as informações, a ser utilizadas nos experimentos, referentes aos preços das ações serão coletadas durante o horário de negociações do Ibovespa, de segunda à sexta, das 10h às 17h, a partir de uma data inicial determinada. Os dados levantados serão de dois tipos, discriminados a seguir.

Fase 2 – Parte 1

O primeiro tipo de dado, refere-se à coleta de informações de 6 sites financeiros diferentes, que trazem notícias sobre a empresa analisada e outras informações relevantes. O objetivo dessa coleta é a classificação dos textos com o uso do processamento de linguagem natural.

Fase 2 – Parte 2

O segundo tipo de dado a ser coletado é o de cotação. Em pequenos períodos tempo – a cada minuto, será registrado o preço da ação de uma empresa ativa no Ibovespa. O objetivo dessa coleta é observar as variações no preço da ação ao longo do pregão. Essa informação será armazenada em um banco de dados e comporá um histórico para a realização da predição do preço da ação.

A empresa escolhida para esta análise é a Companhia de Bebidas das Américas (Ambev), cujo código das ações no Ibovespa é representado por Abev3. A Ambev é uma empresa brasileira de capital aberto, produtora de bens de consumo. Além de ser uma empresa sólida e bem posicionada nos mercados brasileiro e internacional, a coleta de informações será facilitada por ser uma empresa multinacional e com boa reputação.

Fase 3 – Especificação do experimento

Para a coleta dos dados da fase 2, parte 1, foi desenvolvido um programa em *Python* para extração das notícias. Esse programa extrairá as notícias de seis sites financeiros e as disponibilizará para o *Big Data Hadoop* processá-las.

A arquitetura do *Big Data Hadoop* versão 2.7.3 consiste em 01 computador mestre e 01 computador escravo, em ambiente virtualizado com *VMware Workstation* versão 11.1.0 *build-2496824*, em um computador com Windows 10.

Todas as máquinas virtuais estão configuradas com 1 processador, 1 *GByte*s de memória *RAM* e 100 *GByte*s de espaço em disco.

Utilizando a função *Hadoop Streaming*, que permite a utilização de outras linguagens de programação diferentes de *Java*, foi desenvolvido um programa *Map* e *Reduce* em *Python* para a utilização do processamento de linguagem natural, que permitirá a classificação dos textos coletados.

Para a coleta dos dados do item 2, parte 2, foi desenvolvido um programa em *Java* para extração e armazenamento das informações. Os valores dos preços das ações são extraídos de um site financeiro e armazenados em um banco de dados estruturado.

Depois da informação ser coletada no item 2, será possível sua utilização no ambiente de teste, que consiste no computador mestre estar, devidamente, instalado e configurada a aplicação *R-Studio*, possibilitando-se a utilização da linguagem e ambiente de estatísticas *R* em ambiente cliente/servidor.

Com a linguagem e ambiente de estatística *R* configurados e habilitado o pacote *Naive Bayes* E1071, será possível realizar os cálculos matemáticos preditivos com os dados coletados.

O objetivo da utilização do *Big Data*, *R-Studio* e da linguagem e ambiente de estatística *R* é propiciar um ambiente adequado para se coletar, processar, analisar e apresentar as melhores previsões possíveis para um investidor.

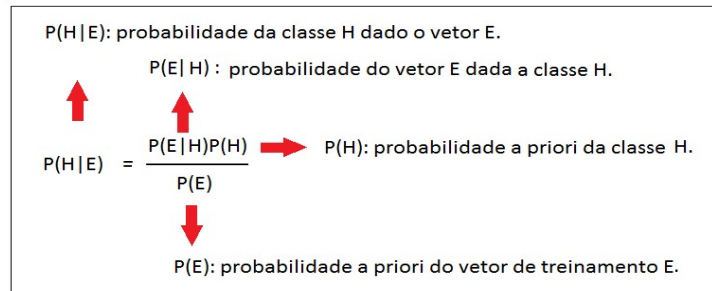
Fase 4 – Execução dos testes

Os testes são divididos em 2 partes.

Fase 4 – Parte 1

A primeira parte desta fase será destinada aos testes com os valores das ações coletados durante o pregão eletrônico. Observando as variações dos valores das ações e aplicando-se o classificador *Naive Bayes*, será possível calcular, no exemplo, a probabilidade do valor das ações da empresa Ambev, em determinado momento do dia.

Fórmula 1 - Classificador *Naive Bayes*.



Fonte: (SANTIDHANYAROJ, 2015) Adaptado e traduzido para o português pelo autor.

Na Fórmula 1 é apresentada a fórmula do classificador *Naïve Bayes* e na Figura 2 o cálculo para identificação da porcentagem de um valor atingido ao longo do pregão. Para o esse cálculo é necessária a construção das tabelas de frequência e probabilidade.

Fig. 2 - Exemplo do cálculo de probabilidade de um valor das ações da empresa Ambev durante o pregão eletrônico.

$$P(\text{Positivo})|18,79) = \frac{P(18,79|) * P(\text{Positivo})}{P(18,79)} = \frac{0,0833 * 0,1}{0,6} = 0,0138833$$

Fonte: Autor.

A função *Naive Bayes* do pacote *E1071* da linguagem e ambiente de estatísticas *R* permite que se calcule, de forma automática, as probabilidades das variações dos preços das ações a partir do momento que os dados são obtidos na parte 2 da Fase 2. Com a diferença das oscilações dos preços das ações da Ambev, é possível estimar o valor que esta ação atingirá ao longo do dia.

O Objetivo da primeira parte do item 4 é acompanhar e prever as oscilações dos valores das ações da Ambev.

Fase 4 – Parte 2

A segunda parte da Fase 4 consiste em classificar os textos coletados na Fase 2 (Parte 1). Para a classificação de textos será utilizada o processamento de linguagem natural.

Os textos coletados serão processados no *Big Data* pelo *Map* e *Reduce* desenvolvido em *Python*. Ao final, os textos serão classificados positiva ou negativamente e aferidos pela fórmula 2.

Fórmula 2 - Fórmula de aferição de preços.

$$\frac{\text{total de palavras}}{\text{total de palavras} \cdot \text{total de documentos}}$$

Fonte: (KIM, 2014), adaptado e traduzido para o português pelo autor.

O resultado dessa fórmula será o valor a ser acrescentado ou subtraído na predição anteriormente calculada, com o objetivo de aproximar-se do valor real da ação naquele momento.

Fase 5 – Análise dos resultados

Nessa seção, são resumidos a hipótese deste trabalho, o método de pesquisa e os resultados da análise obtidos após a realização do experimento. E, ainda, serão analisados os resultados apresentados pelo processo de predição, comparados com o valor das ações durante o pregão eletrônico e analisada a aferição. Caso a aferição do processo preditivo apresente resultados muito diferentes dos valores do pregão eletrônico, serão investigadas as possíveis causas e propostas adequações na aferição dos resultados. Essa seção discute também a possibilidade de estudos futuros e melhoria no processo proposto.

1.5 Organização do trabalho

A seção 2: Introdução Teórica – apresenta a revisão das teorias necessárias para o desenvolvimento desta dissertação como mercado de capitais, probabilidades, teorema de *Bayes*, classificador *Naive Bayes* e *Big Data*.

A seção 3: Coleta dos Dados para o Experimento - apresenta os exemplos utilizados, os tipos de dados coletados para o experimento, a periodicidade da coleta, a extração de dados pelo *Big Data* e a forma de armazenamento e consulta em Banco de Dados.

A seção 3: Execução do Experimento - apresenta os detalhes do ambiente técnico com *Big Data* e *R-Studio* e os detalhamentos das funções matemáticas do *Naive Bayes* para realização da predição e da classificação de texto com o processamento de linguagem natural.

A seção 4: Análise dos Resultados - apresenta o resultado da predição dos preços das ações da Ambev e o compara com os valores durante o pregão eletrônico e os resultados *Big Data* e classificação dos textos. Os resultados são divididos em três partes, sendo a primeira, a predição do preço da ação Ambev com a utilização do classificador *Naive Bayes*; a segunda, a classificação de textos com a PLN e a terceira parte, o resultado final.

Na seção 5: Conclusão - são apresentadas as considerações finais e propostos possíveis trabalhos futuros.

2 ESTADO DA ARTE

2.1 Introdução

Nesta seção, serão apresentados os conceitos relevantes para o trabalho ser desenvolvido e também os trabalhos relacionados. A seção será organizada como a seguir. Na subseção 2.2 uma introdução do mercado financeiro; na subseção 2.3 as diferenças entre análise fundamentalista, análise técnica e predição; na subseção 2.4 a probabilidade e estatística; na subseção 2.5 o Teorema de *Bayes*; na subseção 2.6 o *Big Data*; em 2.7 o *MapReduce*; na subseção 2.8 os trabalhos relacionados e; na subseção 2.9 a conclusão dessa seção.

2.2 Mercado financeiro

Pelo BMF&Bovespa, pode-se comprar e vender ações de empresas participantes. Uma ação é uma parte de uma empresa, a menor parcela do seu capital social. Ao adquirir uma ou mais partes de uma empresa, uma pessoa torna-se sócia dela.

Existem diferentes definições para preço de uma ação, ele pode ser definido como o valor que alguém está disposto a pagar ou o que a última pessoa pagou por algo. Pode ser definido também como sendo a interseção das curvas de oferta e demanda. (ELDER, 2004).

Esta última definição mostra-se mais aplicável ao contexto, pois o preço das ações está relacionado à oferta e à demanda praticadas pelo mercado e não necessariamente pelo último valor pago.

O preço das ações está sujeito às oscilações de mercado, como crises econômicas locais ou globais, fusões e aquisições de empresas, publicações de balanços contábeis, catástrofes naturais ou não e outros fatores. Prever o preço de uma ação é analisar a saúde financeira de uma empresa e acompanhar essas oscilações.

2.3 Análise técnica, análise fundamentalista e predições

São muitas as ofertas de produtos e opções que as instituições bancárias e corretoras oferecem aos seus clientes como forma de diversificarem seus investimentos. Para os investidores que não procuram opções de investimento tradicionais como Comprovante de Depósito Bancário (CDB), Renda Fixa,

Poupança, etc., existem os investimentos de riscos chamados de Rendas Variáveis ou compra e venda de ações.

As compras e vendas de ações são, basicamente, divididas em três tipos, dependendo do perfil de investidor e expectativas de prazos: as de longo prazo, as de curto prazo e os dividendos. Para analisar as ações de longo prazo, podem ser realizadas as chamadas análises fundamentalistas já para as de curto prazo, são realizadas análises técnicas. O terceiro tipo, os dividendos, refere-se ao tipo de investidor que compra ações e não se preocupa com as variações do preço da ação, porque seu interesse é pelo pagamento de dividendos, ou seja, dos lucros da empresa divididos entre os investidores, geralmente pago anualmente ou semestralmente.

A análise fundamentalista utiliza, primariamente, modelos matemáticos que usam uma variedade de fatores como taxas de juros, balanços contábeis, risco-país, projeções macro e microeconômicas para determinar o fluxo de caixa e trazê-lo ao valor presente a fim de projetar o preço alvo de determinada ação. Análise técnica é o estudo da ação do mercado, primariamente, por meio de uso de gráficos, com o objetivo de prever as tendências futuras dos preços. (LEMOS, 2015)

Na análise fundamentalista, um estudo é baseado nas informações existentes até a presente data, sobre a empresa alvo, visando identificar a longo prazo se as ações da empresa sofrerão uma valorização que justifique a aquisição. A análise fundamentalista projeta um preço que as ações podem atingir durante um período de tempo, geralmente meses ou anos.

Existem muitas formas que os especialistas em análise fundamentalista utilizam para projetar os preços das ações, algumas das mais utilizadas são (MOORE 2015):

- P/L (Preço/Lucro): preço da ação no mercado / lucro por ação;
- RPL (Retorno sobre patrimônio líquido): lucro líquido da empresa / patrimônio líquido;
- LPA (Lucro por ação): lucro líquido / quantidade de ações emitidas;
- P/VPA (Preço da ação / valor patrimonial por ação)
- *Dividend Yield*(%): dividendos / preço da ação

- Ebitda: *Earning Before Interest, Taxes, Depreciation and Amortization*
- VE/Ebitda (Valor de empresa sobre Ebitda).

Os preços das ações variam a todo momento. Não é possível prevê-los com absoluta certeza de acertos. Os preços tendem a subir por vários motivos como: tendências de mercado, valorização da empresa, aumento de capital, recebimento de investimento estrangeiro, etc. e tendem a descer quando é publicado um balanço negativo, ou há perda de investimento, escândalos entre outros.

Outro fator que causa a oscilação dos preços das ações é a especulação de mercado. O investidor precisa reconhecer a existência de um fator especulativo em sua carteira de ações e é sua tarefa manter esse componente dentro de limites estreitos, além de estar preparado financeira e psicologicamente para resultados adversos que possam vir, seja em curta ou longa duração. Um investidor calcula o valor de uma ação com base no valor dos negócios a ela relacionados e aposta que uma ação subirá de preço porque alguém pagará ainda mais caro por ela. (GRAHAM, 2015)

As variações dos preços não seguem, fielmente, uma ou mais regras e também não existe um padrão de comportamento que determine que sempre que uma empresa estiver em um determinado momento econômico, uma valorização ou redução de preço se repetirá, como em situações passadas. Uma tentativa de explicar as variações de preços que sofrem as ações, pode, talvez, ser expressa em duas palavras de mercado: oferta e procura.

Não é de desconhecimento do mercado que quando um produto está sendo muito procurado, seu preço tende a subir e quando o produto está em baixa, seu preço tende a baixar. Muitos comerciantes e empresas, sendo grandes ou pequenas empresas, utilizam-se de tal prática causando as variações nos preços.

Um exemplo é o mercado automobilístico, quando um novo carro é lançado por uma montadora, abrindo o mercado em um novo segmento, geralmente, o produto é lançado por um preço alto. A montadora, por estar abrindo o mercado e não ter um preço da concorrência, lança o produto e observa o comportamento dos consumidores. Existindo demanda, o preço tende a subir ainda mais, caso não exista, o preço cairá ou mesmo o produto poderá até ser descontinuado após um

período de tempo. A regulação dos preços é definida pela demanda de mercado, causando variações, muitas vezes, inesperadas.

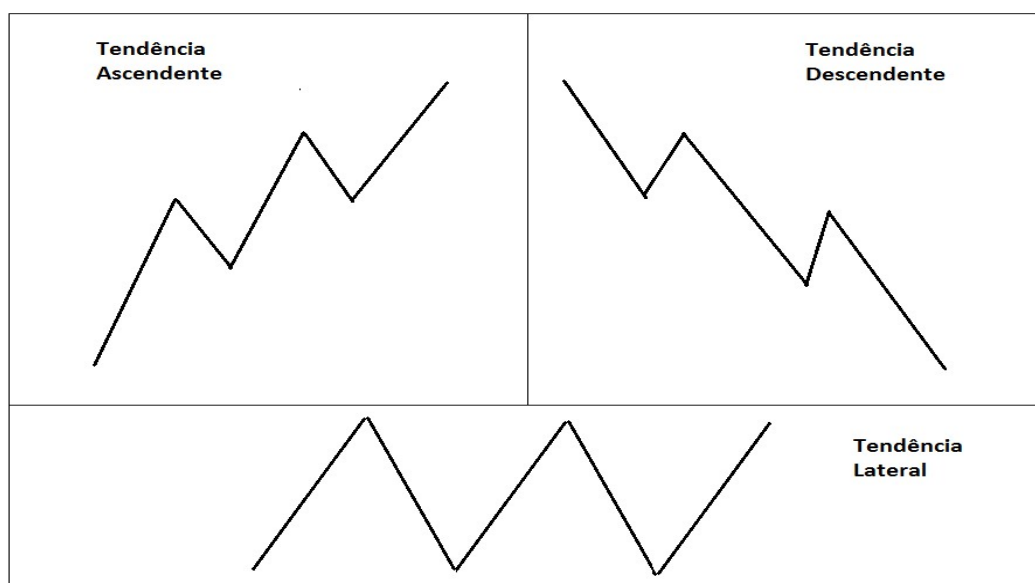
Para o investidor que sempre busca a maximização dos lucros, essas inesperadas oscilações nos preços podem causar preocupações e até mesmo desesperos. As alterações inconstantes nos preços são causadas por ações de grandes empresas e grandes investidores. Tais empresas e investidores compram de milhões a bilhões de ações, causando uma “onda” no mercado. O pequeno investidor, geralmente, não consegue causar variações no mercado.

Na análise fundamentalista, que projeta o preço da ação em um período para frente, não conseguiria levar em consideração todas as variações de preços que a ação poderia sofrer durante o seu período de análise, embora alguns analistas fundamentalistas já utilizem a análise técnica para realizarem seus cálculos. Esta depende, essencialmente, da análise de balanços para identificar a saúde financeira da empresa e os riscos para projetar o preço que a ação poderia alcançar.

Na análise técnica, geralmente na forma gráfica, é possível identificar as variações dos preços e muitas oscilações de mercado que ocorrem a todo momento como notícias, tragédias e acontecimentos extraordinários que afetam o preço da ação naquele momento e que são refletidos, quase que instantaneamente, nas ferramentas gráficas. Também é possível identificar tendências e padrões na forma gráfica.

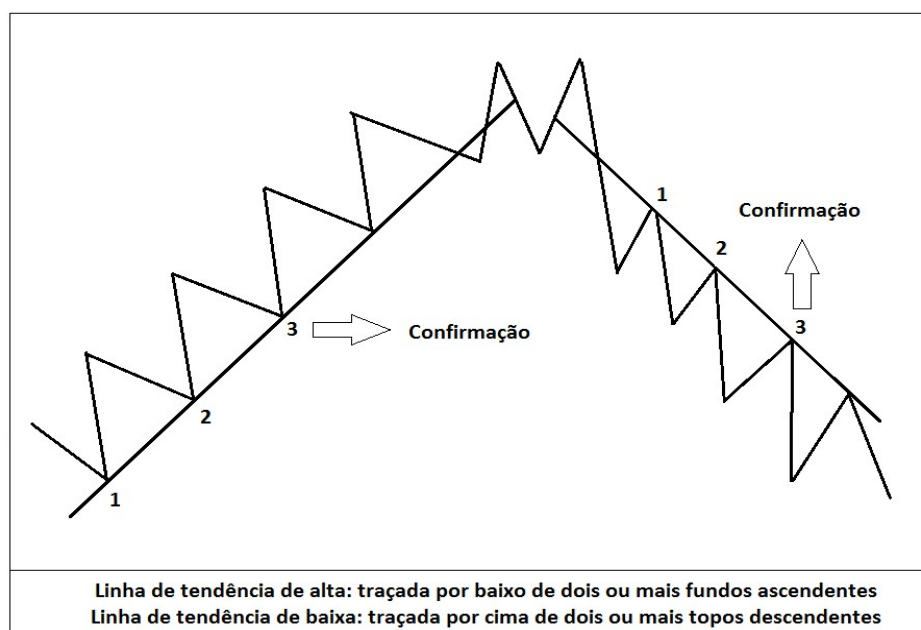
Uma forma simples de se analisar os gráficos, segundo a análise técnica, são as linhas de tendências conforme a Figura 4. Esta linha reta traçada de baixo para cima, unindo no mínimo dois fundos ascendentes proeminentes.

A análise técnica segue um conjunto de regras e ações para determinar o preço das ações, de acordo com ela o mercado se movimenta em tendências.

Fig. 3 - Linhas de tendências de mercado.

Fonte: (LEMOS, 2015). Adaptado pelo autor.

Tendências são tipos de movimentos que indicam para onde o mercado está seguindo. Podem ser para cima, para baixo, ou para o lado. A Figura 5 apresenta os tipos de tendências.

Fig. 4 - Confirmação da tendência de alta e baixa, após a união do terceiro ponto.

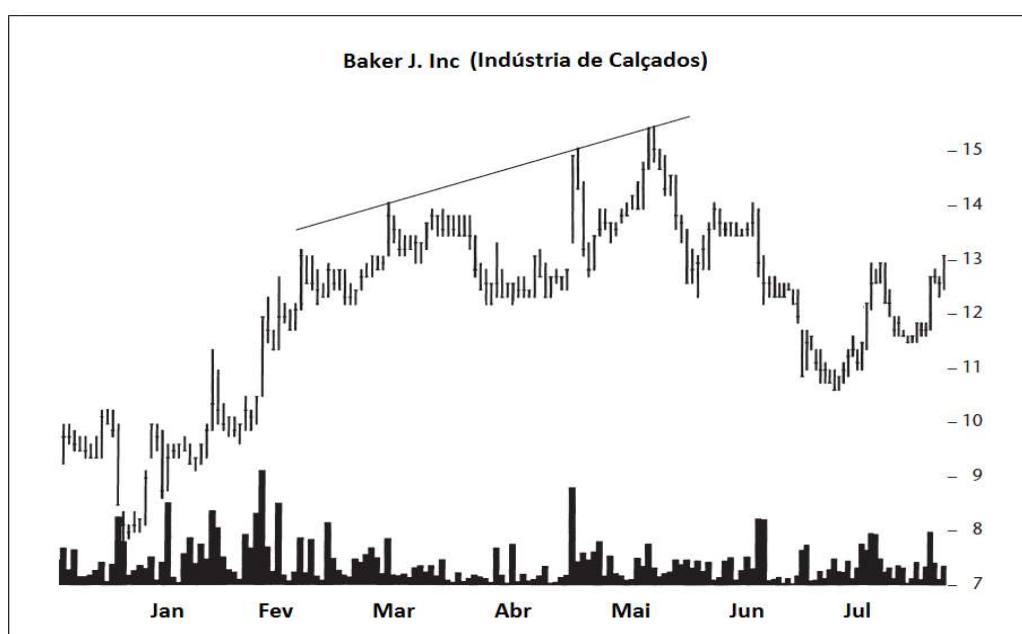
Fonte: (LEMOS, 2015). Adaptado pelo autor.

Segundo (LEMOS, 2015), a união após do terceiro ponto, serve como confirmação da tendência se manter. A tendência se manterá tanto na ascendência como na descendência.

A análise técnica analisa alguns padrões que podem ocorrer em alguns gráficos. A confirmação da união do terceiro ponto não necessariamente deve ser identificada como uma regra e que sempre ocorrerá. É possível que ocorra em algumas situações, mas é um risco para o investidor sempre identificar como fator positivo e que sempre promoverá a subida do preço da ação, ou mesmo acreditar que seu preço descerá mesmo com a união dos três pontos.

No gráfico da Figura 5 é apresentado o valor da empresa de calçados americano *Baker J. Inc.*. Nos meses de março, abril e maio é possível visualizar uma reta de tendência ascendente unindo os três pontos. Nessa situação, com a união dos três pontos, a tendência seria mantida em alta e os preços continuariam a aumentar. No mês de junho, a tendência se inverteu, fazendo com que os preços se desvalorizassem. Em julho, os preços diminuíram ainda mais.

Fig. 5 - Análise de tendência. União do terceiro ponto e o preço das ações não se mantiveram em alta.



Fonte: (BULKOWSKY, 2005). Adaptado e traduzido pelo autor.

Se um investidor tivesse seguido a tendência de alta e adquirido mais ações é certo que sofreria perdas financeiras.

Uma linha de tendência pode não representar exatamente o que acontecerá. As linhas de tendência ajudam a identificar as subidas e descidas, porém não representam previsões baseadas na matemática. No gráfico, as linhas de tendência estão ligando os pontos mais altos ou mais baixos no gráfico em uma linha reta.

Uma linha de tendência, mesmo unindo os três pontos em uma reta ascendente, pode indicar que a tendência dos preços é subir, mas, na prática, isso pode não se efetivar, necessariamente. Diversos eventos externos que podem afetar os preço da ação não foram computados, como tragédias, publicações de balanços, escândalos, etc. Nesse cenário o investidor só perceberia que algo aconteceu quando visualizasse o gráfico das ações caindo e quando mais nada poderia fazer, preventivamente.

A análise técnica no contexto desses exemplos, não está realizando previsões e sim previsões e tendências na forma de gráficos. A previsão deste trabalho, utilizará o teorema de *Bayes* específico para cálculos de previsões. Com as previsões, é possível mostrar ao investidor a previsão do dia, até mesmo antes do pregão eletrônico iniciar. Possibilitando uma visão de investimento e hipóteses de melhor momento para compra e venda das ações. Tendo em vista os diversos eventos que ocorrem pelo mundo e que podem influenciar o preço da ação alvo, o *Big Data* pode contribuir para ajustar seu preço em tempo real, no momento em que as informações chegam à Internet. Dessa forma, a previsão realizada anteriormente é alterada para uma outra previsão mais próxima da realidade.

2.4 Probabilidades e estatísticas

A estatística é uma parte da matemática que fornece métodos para a coleta, organização, descrição, análise e interpretação de dados que viabilizam a tomada de decisões. Probabilidade são modelos matemáticos que explicam os fenômenos estudados pela estatística em condições normais de experimentação. (CORREA, 2003)

Pela probabilidade e estatística é possível utilizar um modelo matemático que pode antecipar os valores dos preços das ações de uma determinada empresa participante no BMF&Bovespa a serem alcançados, em determinado período. Essa

predição auxilia a pessoa ou as empresas a tomarem decisões de compras ou vendas de ações baseando-se em probabilidades.

Um modelo muito utilizado é o modelo *Bayesiano* que se baseia no teorema de *Bayes*, de *Thomas Bayes*.

Thomas Bayes (1701-1761) viveu na Inglaterra, estudou teologia na Escócia, publicou o livro *A doutrina dos fluxions*. Com base nesse livro, *Bayes* foi eleito, em 1752, membro da Real Sociedade, entidade britânica criada em 1645. Após sua morte, um amigo apresentou à Real Sociedade um artigo encontrado entre suas anotações com o nome Ensaio, em que o autor buscava resolver um problema na doutrina das probabilidades. Nesse artigo estava demonstrado o famoso teorema de *Bayes*. Após a publicação do artigo, o trabalho caiu no esquecimento e foi resgatado anos mais tarde pelo matemático *Pierre-Simon de Laplace* (1749-1827) que o revelou ao mundo. (PENA, 2016).

O teorema de *Bayes* tem sido utilizado em áreas como finanças, biologia, seguradoras de veículos, sistemas que evitam fraudes e outras, nas quais se deseja realizar previsões probabilísticas.

2.5 Teorema de *Bayes*

Naive Bayes (NB) é um classificador probabilístico que utiliza as regras de *Bayes*.

A Fórmula 3, representa a regra *Bayesiana*,

Onde:

c = classe

d = elemento sob consideração.

Foi decomposto o termo $P(d|c)$ em classificadores independentes de acordo com a regra Bayesiana. (SANTIDHANYAROJP., 2015).

Fórmula 3 - Classificador Naive Bayes.

$$P(c | d) = \frac{P(c)P(d | c)}{P(d)}$$

Fonte: (SANTIDHANYAROJP., 2015).

Naive Bayes são classificadores probabilísticos simples, amplamente utilizados como modelos de treinamentos e resultados empíricos. (NUNZIO, 2012)

É possível melhorar a predição das hipóteses com o treinamento dos classificadores NB.

2.6 *Big Data*

Atualmente, um alto volume de dados como: registros de navegações, textos e documentos, transações comerciais, registros bancários, gráficos financeiros, imagens médicas, vídeos de vigilância, marketing, telecomunicações e dados de mídias sociais pode ser facilmente coletado ou gerado de diferentes fontes, em diferentes formatos em distintas aplicações, em tempo real, na era do *Big Data*. (LEUNG, 2014).

Big Data refere-se aos grandes e complexos conjuntos de dados com que as tradicionais ferramentas de processamento de dados não conseguem lidar.(CHAN, 2013).

Uma aplicação de código aberto bastante conhecida de *Big Data* é o *Hadoop* mantido pela *Apache Software Foundation*. É possível construir diversas arquiteturas de *Big Data* com *Hadoop*, para atender a diferentes propósitos.

Apache Hadoop é uma coleção de bibliotecas de softwares que permite um processamento distribuído em larga escala de um número grande de computadores, conhecido como *clusters*, usando uma programação de modelos simples. Sua arquitetura escalonável permite a utilização de um simples servidor ou de centenas de servidores, oferecendo processamento, armazenamento e alta disponibilidade. (APACHE HADOOP, 2016). Fazem parte os seguintes módulos:

- a) *Hadoop Common*: Módulo comum que dá suporte a outros módulos.
- b) *Hadoop Distributed File System (HDFS)*: um sistema distribuído que provê altas taxas de transferências dos dados de aplicação.
- c) *Hadoop Yarn*: uma estrutura que provê agendamento de processos e gerenciamento de recursos.
- d) *Hadoop MapReduce*: um sistema baseado no *Hadoop Yarn* para processamento paralelo em larga escala.

2.7 Map Reduce

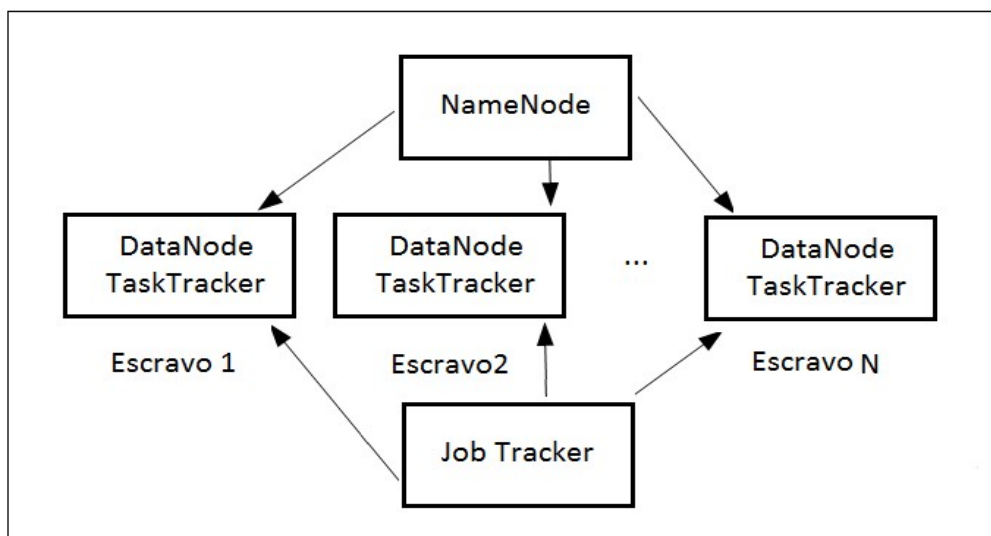
MapReduce é um modelo de programação para processamento de dados de forma paralela. O modelo de programação *MapReduce* consiste em duas funções, *map()* e *reduce()*. A função *map()* recebe a entrada dos valores/chaves e produz uma lista intermediária. O *MapReduce* processa a lista intermediária e os valores/chaves, e os encaminha para a função *reduce()*, produzindo o resultado final. (MANIKANDAN, 2015).

Os componentes do *MapReduce* são:

- a) *Name Node*: gerencia os dados armazenados no *Hadoop Distributed File System*.
- b) *Data Node*: armazena os blocos de informações nos computadores escravos.
- c) *Job Tracker*: agenda, aloca e monitora a execução dos serviços nos computadores escravos.
- d) *Task Tracker*: executa as operações do *Map* e *Reduce* e informa o progresso das atividades.

Na Figura 8 é apresentada a estrutura do *MapReduce* em funcionamento. Pode-se ter dois, ou mais, computadores escravos sendo gerenciados pelo *Job Tracker*. Cada computador escravo possui o processo *TaskTracker*, responsável por dividir as tarefas em cada máquina.

Fig. 6 - MapReduce processando com o computador Mestre e Escravos.



Fonte: (MANIKANDAN, 2015). Adaptado e traduzido para o português pelo autor.

Com essa arquitetura, o *Map* e *Reduce* fornece escalabilidade, tolerância a falhas e performance.

2.8 Trabalhos relacionados

Os trabalhos apresentados nesta seção servirão de base para o desenvolvimento de pesquisa deste trabalho.

Tabela 1 - Trabalhos relacionados.

	SHIHAVUDDIN,2010	GHOFRANI,2015	RAO,2015	Kim,2014	Este trabalho
<i>Big Data</i>			x		x
<i>Naive Bayes</i>	x	x			x
Proce.Ling.Natural				x	x
Predição de Preços	x				x

Fonte: Criado pelo autor

A tabela 1 apresenta os trabalhos de pesquisa relacionados.

O classificador NB possibilita a um investidor do mercado financeiro realizar previsões e conhecer, antecipadamente, as variações positivas e negativas dos valores das ações e direcionar suas estratégias de negócios otimizando lucros, mitigando riscos e melhorando as estratégias financeiras.

O classificador NB foi utilizado para realizar a predição de valores das ações *FTSE100*. O *FTSE100* é um índice calculado pela *FTSE The Index Company*. É composta por ações de 100 empresas na Bolsa de Valores de Londres. Foram coletados 170 artigos da Internet com notícias relevantes sobre as principais empresas que poderiam afetar o índice *FTSE100*. Para treinar os classificadores, foram utilizados 140 fluxos de dados. (SHIHAVUDDIN, 2010)

Nos 170 artigos, usou-se o classificador de texto para realização da predição. Os artigos foram extraídos do site financeiro <http://www.thisismoney.co.uk>. Uma das dificuldades, segundo o autor, foi o de decidir qual o percentual de influência extraído dos artigos poderia ser aplicado para alterar o índice *FTSE100*. Depois das coletas das notícias, o classificador NB calculava uma probabilidade de influência no índice *FTSE100* com as notícias do dia e aplicava na predição do índice do dia seguinte. Com esse processo sendo repetido durante 60 dias, conseguiu-se uma predição de precisão de 58.62%, 66.62% e 79.4% em três rodadas de testes.

A tabela 2 apresenta uma amostra das coletas das palavras e como foram classificadas. A quantidade de ocorrências nos textos coletados determina a sua probabilidade. As palavras foram contadas, independentemente, de sua posição no texto. A probabilidade foi calculada em relação ao número total de ocorrências das palavras no texto.

Tabela 2 - Exemplo de classificação de textos atribuídos a baixa e alta probabilidade para composição da predição do índice *FTSE100*.

Índice de Vocabulário	Palavras	Ocorrências das palavras	Probabilidade
2.00	compartilhado	490	0.0112
3.00	grande	56	0.0009
4.00	óleo	125	0.0035
5.00	exploradores	7	0.0002

Fonte: (SHIHAVUDDIN, 2010), adaptado e traduzido para o português pelo autor.

O autor apresentou uma importante questão nesse artigo que é o percentual de influência das palavras aplicadas na predição. No experimento, não foi explicado como esse cálculo foi realizado.

Outro ponto, foi a utilização de somente um site de notícias financeiras como fonte de dados. Poderiam ter sido usadas mais fontes de dados, diferentes sites de notícias, consultas a banco de dados, pesquisas de mercado e outros, com um número maior de informações de fontes variadas é possível se obter maior aferição na predição.

Os dados extraídos e analisados somente eram comparados ao índice *FTSE100* no dia seguinte, o que impossibilita sua utilização em mercados *Day Trade*, onde se compra e se vende ações no mesmo dia.

Em outro experimento realizado por (Ghofrani, 2015), o classificador *Naive Bayes* foi comparado a outros classificadores como *K-Nearest Neighbor* (KNN), *Support Vector Machine* (SVM). Os três classificadores foram submetidos a testes de tráfego de dados na Internet, classificando textos em diferentes tipos de serviços. Os seguintes serviços foram analisados: *http*, *smtp*, *ftp*, *pop3*, *nnntp*, *msn*, *edonkey*, *ssh*, *bittorrent* and *https*. Para o experimento foram utilizados o total de 13536 fluxos de dados com 50 fluxos de classes de treinos para os 3 classificadores. O classificador *Naive Bayes* foi o que, na média dos serviços analisados no experimento, conseguiu o melhor resultado em termos de precisão.

A Tabela 3 apresenta um comparativo de eficiência individual entre os classificadores KNN, SVM e *Naive Bayes* em um experimento de 13536 fluxos de dados.

Tabela 3 - Comparativo de eficiência individual dos classificadores KNN, SVM e Naive Bayes.

Serviços	KNN	SVM	Naive Bayes
http	90,9	99,85	94,1
smtp	97	85,95	91,95
ftp	96,95	87,05	95,4
pop3	90,35	78,3	97,35
nntp	98,2	97,95	97,05
msn	98,36	80,6	98,36
edonkey	93,8	66,34	91,58
ssh	93,5	72,4	98,37
bittorrent	92,15	84,9	96,25
https	80,6	59,91	89,65
Média	93,18	81,32	94,4

Fonte: (Ghofrani, 2015), adaptado e traduzido para o português pelo autor.

A média de resultados foram 93.18% para o *K-Nearest Neighbor*, 81.32% para *Support Vector Machine* e 94,4% de precisão para o *Naive Bayes*. O autor não deixou claro como os testes para esses serviços de Internet foram realizados. Para o protocolo *HTTP*, o NB ficou em segundo lugar com nota 94,1 enquanto para *POP3* ficou em primeiro com nota 97,35.

Outra dúvida foi a quantidade de treinos para os classificadores com que o autor iniciou os experimentos. Ele começou com a quantidade de 50 vezes definida, normalmente o número de treinos dos classificadores varia de acordo com o universo analisado e as expectativas a serem alcançadas.

Assim como o NB, o KNN e o SVM são bastante utilizados e o autor conduziu bem o experimento, explorando os classificadores nos principais serviços de Internet em que são comumente utilizados.

Em um outro artigo realizado por (RAO, 2015), o autor comenta sobre uma importante questão que é a qualidade da informação utilizada pelo Big Data.

Com o advento do *WWW* no início da década de 90, foi acrescentada uma nova dimensão para a qualidade de dados e o *HTML* foi desenhado para marcação de dados. As informações que alimentam o *Big Data* podem ser de forma estruturada e não estrutura. A quantidade de dados disponíveis para coleta, muitas vezes, prejudica a qualidade de uma informação, assim como sua veracidade.

Muitos dos dados são coletados por redes de sensores, câmeras *giga pixel*, Internet das coisas, dispositivos de vigilância e monitoramento e media social por múltiplos fornecedores. Os dados precisam serem limpos para serem utilizados. (RAO, 2015).

Nesse artigo o autor comenta que muitas informações de diferentes fontes de dados podem prejudicar a qualidade de uma informação que seria coletada pelo *Big Data*. Diferentes fontes podem apresentar, por exemplo, uma mesma informação com a mesma qualidade, desde que respeitada a veracidade da informação. O autor não comenta as formas utilizadas para a coleta, pois existem diferentes fornecedores de tecnologias e diversas formas de coletas, e o HTML é a estrutura básica das páginas de Internet que podem ser utilizadas para a obtenção de informações, de forma dinâmica.

Em outro experimento realizado por (Kim, 2014), foi utilizado a mineração de textos e análise de notícias para predição do mercado financeiro. Foram coletados 78216 notícias de artigos financeiros de duas empresa M e H, durante o período de um ano, em 2011. Foi utilizado o processamento de linguagem natural e também criado um dicionário de palavras de sentimentos. Esse dicionário compreende palavras de sentimentos positivos e negativos para compor a predição.

Foram removidas as palavras de paradas, números, palavras em inglês, codificações *tag html* para a processamento de processamento natural funcionar.

Segundo o autor, com essa arquitetura conseguiu-se da empresa M 65.2% e da empresa H 60.3% de acertos de predição do mercado financeiro. As diferenças de porcentual das empresas ocorreram por contextos diferentes de cada empresa.

O autor, de origem Koreana, realizou o experimento em seu idioma, o que poderia causar um comportamento diferente na utilização da linguagem de processamento natural em outras línguas, como o inglês e português, por exemplo. Segundo o próprio autor, a linguagem koreana possui muitos contextos similares e homônimos o que poderia causar variação no porcentual de acertos se comparado a outros idiomas.

Também não foi explicado o método para verificação dos erros e acertos das análises dos sentimentos e opiniões ou mesmo se foi realizado de forma automática ou manual. O autor conduziu bem os experimentos, além de conseguir um porcentual alto de acertos.

2.9 Conclusões da seção

O Teorema de *Bayes* é fundamental para análise das probabilidades. O classificador *Naive Bayes* pode ser utilizado para realizar previsões financeiras.

Naive Bayes teve uma média de acertos melhor se comparados a classificadores como KNN e SVM em um experimento de teste de fluxo de dados pela Internet.

O dicionário de dados se fez necessário para o processamento de linguagem natural identificar termos positivos e negativos na análise de sentimentos.

Existe carência de ferramentas de previsão de preços das ações no mercado financeiro com a utilização do classificador *Naive Bayes* e *Big Data*.

3 DESENVOLVIMENTO

Esta seção descreve as principais atividades de desenvolvimento e implantação necessárias para a operacionalização do ambiente de testes.

3.1 Ambiente *Big Data*

A estrutura de *Big Data* desenvolvida para os experimentos de predição dos preços das ações, tem como objetivo a obtenção de informações de vários tipos coletadas da Internet, sobre a empresa Ambev, objeto desse estudo. Notícias financeiras, de aumento ou diminuição de vendas, publicações de balanços financeiros, tragédias entre outros que, de alguma forma, podem influenciar os valores dos preços das ações para cima ou para baixo.

Um programa extrator coletará as notícias em sites financeiros e irá processá-los no *Big Data* com o uso do processamento de linguagem natural. Um outro programa fará a coleta dos preços das ações, o que ocorrerá durante o pregão eletrônico do Ibovespa de minuto a minuto, de segunda à sexta-feira das 10h às 17h.

3.2 Arquitetura *Big Data*

Para o experimento, foi criada uma arquitetura em um ambiente virtualizado com 01 computador principal denominado Mestre e 01 computador secundário denominado Escravo.

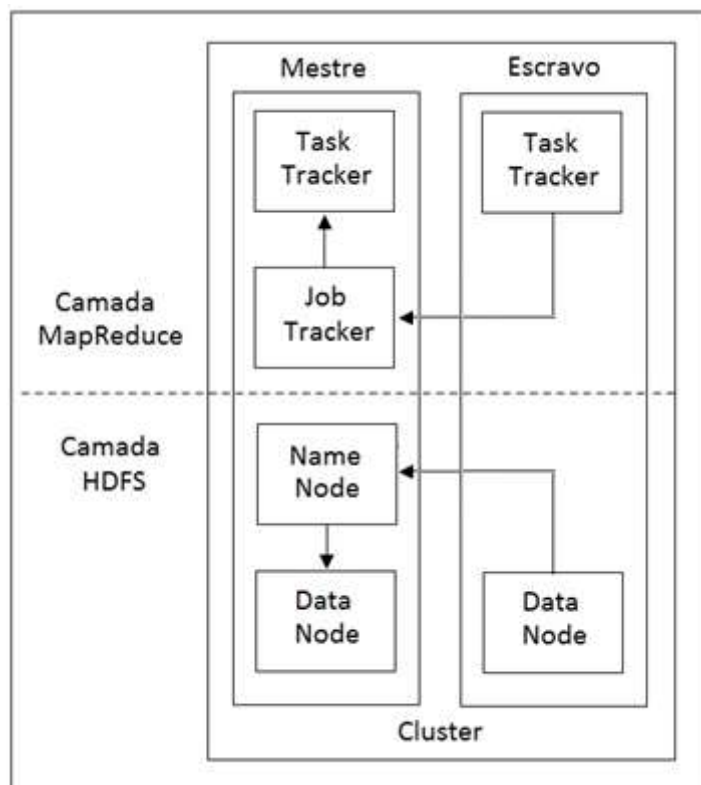
As duas máquinas virtuais são instalações completas do *Linux Ubuntu Server*, versão 15.10 em um ambiente virtualizado com *VMware Workstation Version 11.1.0* build-2496824. É utilizada a versão do *Apache Hadoop 2.7.3* e *Java* versão 1.8.0_91.

Em cada uma das 02 máquinas no *cluster*, é executada uma instância do *DataNode* e *TaskTracker*. O *DataNode* armazena os blocos de informações no computador escravo. O *Task Tracker* é responsável por executar as operações do *Map* e *Reduce*. O *Job Tracker* monitora a execução dos serviços no computador escravo.

A existência de um único *NameNode* no cluster, simplifica a arquitetura do sistema. O *NameNode* é o árbitro e repositório para todos os *metadados HDFS*. O

sistema é concebido de tal maneira que os dados utilizados não fluem através da *NameNode*. (APACHE HADOOP, 2016).

Fig. 7 - Arquitetura do *Big Data HDFS* e *MapReduce* com Mestre e 01 Escravo.



Fonte: (PRAJAPATI, 2013). Adaptado e traduzido para o português pelo autor.

Com uma arquitetura similar a figura 7, a arquitetura com 01 mestre e 01 escravo permitirá a realização dos testes com o processamento de linguagem natural e classificação de textos.

3.3 Programação com *Map* e *Reduce* e Processamento de Linguagem Natural

A arquitetura apresentada no item 3.2 é uma arquitetura em *cluster* e somente o *Big Data* não tem utilidade prática. Para o *Big Data* tornar-se produtivo e atender a algum propósito é necessário o desenvolvimento de programas que atendam aos requisitos da plataforma *Hadoop*.

Nesse contexto, o *Big Data Hadoop* será utilizado para processamento e classificação de textos por meio do processamento de linguagem natural (PLN).

O processamento de linguagem natural está focado na interação entre linguagem natural e computador. É um dos principais componentes da Inteligência Artificial e linguística computacional. Ele fornece uma interação entre computadores e seres humanos e dá aos computadores a capacidade de compreender o discurso humano, com a ajuda do aprendizado de máquinas. (CHOPRA, 2016).

É necessário definir o idioma em que os textos serão classificados. Foi definido o idioma português do Brasil para as funções como lematização e palavras de paradas.

Os programas desenvolvidos em *Java*, para serem executados em ambiente *Big Data* devem estender as classes *Map* e *Reduce*. Essas classes, que são específicas do *Apache Hadoop*, permitem que o programa utilize a computação paralela quando executado. Uma outra forma de executar programas escritos em outra linguagem de programação como, por exemplo, o *Python*, mas com os benefícios da computação paralela será utilizada a função *Hadoop Streaming* nativo em todas as versões do *Hadoop*.

Os programas desenvolvidos em *Python* para processamento de linguagem natural devem utilizar o Kit de Desenvolvimento para Linguagem Natural.

Natural Language Tool Kit (NLTK) é uma extensa biblioteca Python para linguagem natural (PERKINS, 2010).

A versão do *Python* utilizada é a 2.7.12.

Para o desenvolvimento de um *Map* e *Reduce* em Python, são criados outros *Map* e *Reduce* com a extensão *PY*, ficando *Map.py* e *Reduce.py*. Para estes funcionarem no *Hadoop*, basicamente é utilizado o *Stdin* e *Stdout* do *Python*. O *Stdin* é usado para entrada do interpretador e *Stdout* é usado para a saída de instruções de impressão.

Uma forma rápida para testes do *Map* e *Reduce* e que pode-se economizar tempo, pois não é necessário copiar e processar os arquivos no *Hadoop*, é com a execução do comando em Linux:

```
cat texto_a_analisar | python /usr/local/scripts/map.py | sort -k1,1 | python /usr/local/scripts/reduce.py
```

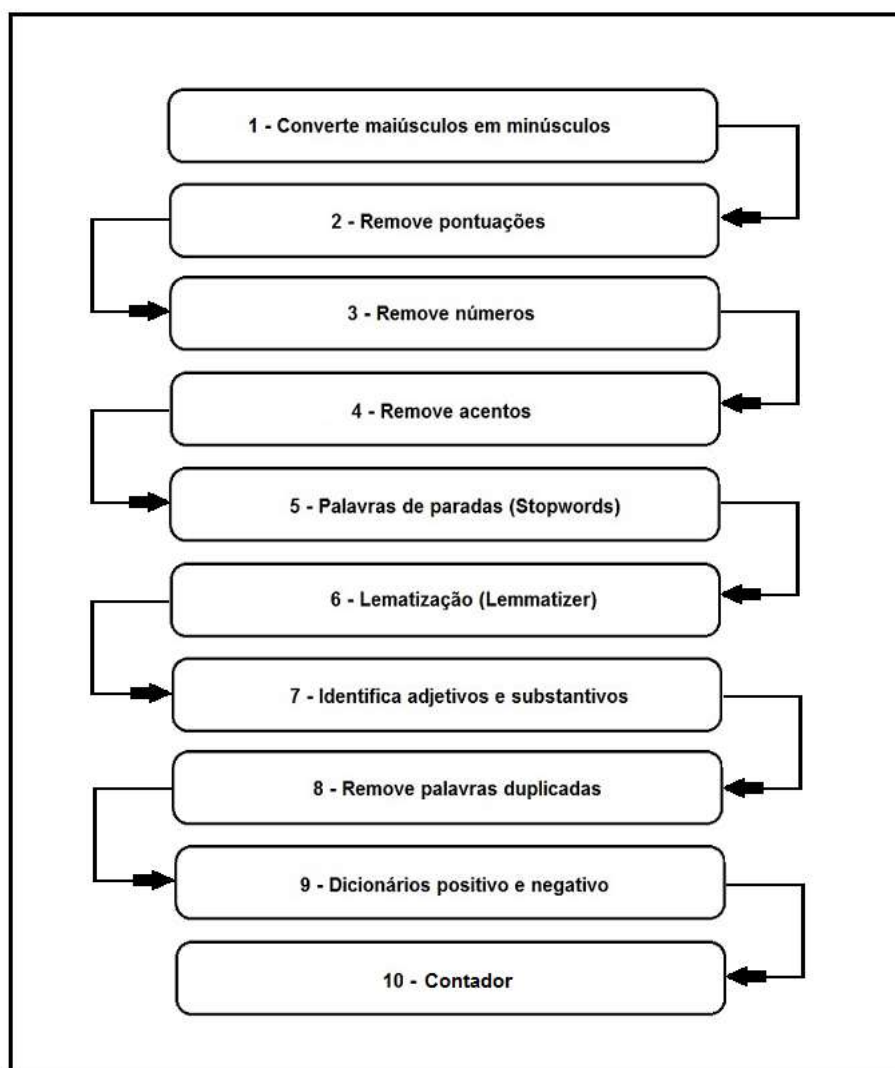
Com esse comando é possível testar *Map* e *Reduce* com *Python* em ambiente *Linux* de forma independente do *Hadoop*.

O objetivo do uso do processamento de linguagem natural é extrair dos textos analisados e processados no *Big Data* os “sentimentos” das informações financeiras que indiquem se os valores das ações subirão, descerão ou se manterão.

Para atingir o objetivo, os textos processados no *Big Data* e analisados com o processamento de linguagem natural passam por 10 etapas. As etapas são apresentadas na figura 8:

Fig. 8 - Etapas do processo de classificação de textos com PLN.

Fonte: Criado pelo autor.



Etapa 1: Converte maiúsculos em minúsculos - todas as palavras que contenham letras maiúsculas são convertidas para minúsculas.

Etapa 2: Remove pontuações - todas as pontuações dos textos são removidas.

Etapa 3: Remove números – todos os números dos textos são removidos.

Etapa 4: Remove acentos – todos os acentos dos textos são removidos.

Etapa 5: Palavras de paradas (*Stopwords*). *Stopword* são palavras que geralmente não contribuem para o significado de uma frase e são removidas. Exemplos de *stopwords* em português: ['a', 'ao', 'aos', 'aquela', 'aquelas', 'aquele', 'aqueles', 'aquilo', 'as']. Todas as palavras dos textos passam pelo *stopwords*.

Etapa 6: Lematização (*Lemmatizer*) – *Lemmatizer* é a conversão de uma palavra para sua palavra raiz. Exemplo:

trabalhando -> trabalho

cachorros -> cachorro

igrejas -> igreja

Todas as palavras dos textos são lematizadas.

Etapa 7: identificação de adjetivos e substantivos – é necessária a identificação dos adjetivos e substantivos nas frases que compõem os dicionários de palavras positivas e negativas. Os adjetivos são identificados por JJ e substantivos por NN.

Etapa 8: remove palavras duplicadas – todas as palavras duplicadas são removidas.

Etapa 9: dicionários de palavras positivo e negativo – conjunto de palavras-chaves com “sentimentos” negativos e positivos identificados em uma lista, indicando se o

valor da ação subirá ou descera. Um pouco abaixo da Figura 12, encontram-se os dicionários positivo e negativo utilizados.

Etapa 10: contador de palavras – são contadas, de forma automática, todas as palavras positivas, negativas, a quantidade de textos analisados e a quantidade total de palavras. Se o total de palavras positivas for maior do que o número de palavras negativas, então o “sentimento do texto” será positivo, caso contrário, negativo. Caso o número de palavras positivas seja igual ao número de palavras negativas, o “sentimento” é neutro e não contabilizado.

Para se executar o *Map.py* e *Reduce.py* com *Hadoop Streaming* foi utilizado o parâmetro abaixo e inserido em um *script*.

```
./hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -  
mapper /usr/local/scripts/map.py -reducer /usr/local/scripts/reduce.py -input  
/user/andre/arquivo01 -output output
```

O resultado do processamento de linguagem natural é baseado em um dicionário de palavras positivas e negativas que ajuda a definir o “sentimento” do texto como positivo ou negativo. Para isso, foram criados os dicionários positivo e negativo que contêm palavras-chaves. Estas palavras-chaves originam-se do estudo dos textos financeiros coletados durante o experimento e que identificam os “sentimentos”. As palavras que compõem os dicionários positivo e negativo estão sem acentos, propositalmente, representando a forma final tratada pelo processamento de linguagem natural.

Dicionário positivo: alta, aumenta, maximo, muito, positivo, valorizacao, abri, ascencao, ganha, dispara, forte, progride, acelera, entra, mais, maior, melhor, sobe, estavel, soma, acredita, alegre, certo, abertura, progressivo, comemora, acerto, inclui, celebra, acumula, lidera, comemora, desconfia, beneficia, bem, alonga, entrada, acima, levanta, salta, pula, avanca, aumento, bom

Dicionário negativo: baixa, diminui, mínimo, pouco, negativo, desvalorizacao, fecha, queda, perde, despenca, fraco, recua, freia, sai, menos, menor, pior, desce, instavel, subtrai, desconfia, triste, incerto, falencia, regressivo, chora, erro, retira, entristece, escasso, atras, lamenta, acredita, prejudica, mal, encurta, saida, abaixo, cai, afunda, rebaixa, retrocede, desconto, ruim

O resultado do contador é salvo em um arquivo texto, ficando disponível para a linguagem e ambiente de estatísticas R inserir a condição positiva ou negativa à predição.

O *Big Data Hadoop* não é eficiente para processar muitos arquivos de pequenos tamanhos como por exemplo arquivos de 5, 10, 15 *kbytes*. No experimento foi medido o tempo para processamento de 100 arquivos pequenos de até 15 *kbytes*, inseridos no HDFS. Foram processados 7 vezes, realizando o processo completo do processamento de linguagem natural e comparando com esses mesmos 100 arquivos concatenados com a função merge do *Hadoop*, gerando um arquivo único.

Tabela 4 - Comparativo de tempo de processamento de 100 arquivos pequenos e 1 arquivo único.

Medição Núm.	Arquivo único (minutos)	100 arquivos (minutos)
1	1m6.592s	1m31.895s
2	1m2.301s	1m38.814s
3	1m3.233s	1m28.939s
4	1m3.130s	1m28.939s
5	1m3.168s	1m29.017s
6	1m4.203s	1m34.044s
7	1m4.206s	1m29.255s
Média	1m3.8s	1m30.5s

Fonte: Criado pelo autor

No comparativo, o *Big Data Hadoop* processou o arquivo concatenado, em média, 27 segundos mais rápido que os 100 arquivos separados, com uma média de 1 minuto e três segundos. Já os arquivos separados demoraram, em média, 1 minuto e trinta segundos, sendo identificado um melhor tempo para um único arquivo concatenado.

3.4 Programa coletor de dados de sites

Para a realização da predição dos valores dos preços das ações durante o pregão eletrônico, será necessária a coleta de informações extraídas de sites financeiros. As informações coletadas comporão um valor a ser calculado por uma fórmula para ser somada ou subtraída à predição dos preços das ações da Ambev.

Os sites que serão analisados são os seguintes:

- 1) <http://aovivo.folha.uol.com.br/2016/01/05/4536-aovivo.shtml?cmpid=mercdir>
- 2) <http://g1.globo.com/tudo-sobre/ambev>
- 3) <http://www.infomoney.com.br/ambevsa/noticias>
- 4) <http://busca.r7.com/?q=ambev&media=>
- 5) <https://buscador.terra.com.br/Default.aspx?source=Search&ca=l&query=ambev>
- 6) http://www.valor.com.br/search/apachesolr_search/%22AmBev%22?solrsort=created%20desc

Os seis sites apresentam informações sobre a empresa Ambev e estão, constantemente, sendo atualizados por terceiros sempre que uma nova notícia é divulgada. As notícias são publicadas nos sites na forma de texto corrido. A coleta de informações não foca a captura do valor do preço das ações ao longo do dia e sim informações relevantes que ajudarão a entender e mensurar as possíveis variações do preço das ações, pela análise do processamento de linguagem natural.

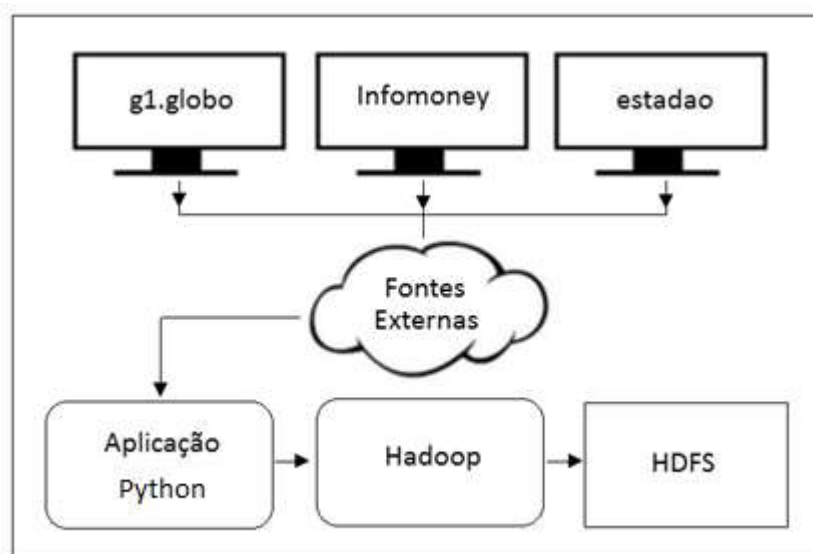
Nos sites de notícias financeiras e de empresas específicas da bolsa de valores, geralmente, são disponibilizadas várias informações na mesma página, como indicadores e notícias de outras empresas, propagandas, fotos, hipertextos para outras páginas e muitas outras informações que não são importantes para o estudo. Esse excesso de informações será descartado no momento da análise dos textos.

A coleta das informações dos sites será realizada por um programa desenvolvido em *Python*, denominado coletor. Para cada site, existe um coletor programado, com endereço específico do site, gerando uma saída única.

O coletor realiza a leitura dos sites, sem a necessidade de abri-los como página *Web*. Para o desenvolvimento do coletor, foi utilizado a biblioteca *Beautiful Soap* do *Python*. Essa biblioteca permite extrair dados de arquivos *HTML* e *XML* e gerar uma saída texto.

O arquivo texto é copiado para o *Hadoop* processando-o em modo cluster pelo *Map.py* e *Reduce.py* apresentados na seção 3.3.

Fig. 9 - Coleta de informações de sites para composição de aferição de predição dos preços da ação da Ambev.



Fonte: Criado pelo autor.

A Figura 11 apresenta a coleta de informações dos seis sites de notícias da Ambev até a informação chegar no *HDFS*. Na sessão 3.5 é apresentado o cálculo de ajuste.

3.5 *RStudio* para cálculo de aferição

O *RStudio* mais a linguagem e ambiente de estatísticas R com suas inúmeras bibliotecas próprias e de terceiros torna-se uma ferramenta importante para análise dos dados. Ele é um ambiente de desenvolvimento integrado (IDE) para R, que inclui um console, editor de sintaxe, que suporta a execução direto do código, bem como ferramentas gráficas, históricos, depurações e gerenciamento de espaço de trabalho. (RSTUDIO, 2016).

Para a realização do cálculo de aferição, os textos classificados pelo processamento de linguagem natural serão identificados como positivos, negativos e neutros.

Serão positivos quando o resultado do processamento da linguagem natural retornar como positivo e negativos, quando o resultado retornar negativo e neutros quando a quantidade de palavras identificadas como positivas forem iguais à quantidade de palavras negativas.

Após esses possíveis três resultados apresentados pelo processamento de linguagem natural, será utilizada a Fórmula 2 de aferição.

Fórmula 4 - Fórmula de aferição de preços.

$$\frac{\text{total de palavras}}{\text{total de palavras} \cdot \text{total de documentos}}$$

Fonte: (KIM, 2014), adaptado e traduzido para o português pelo autor.

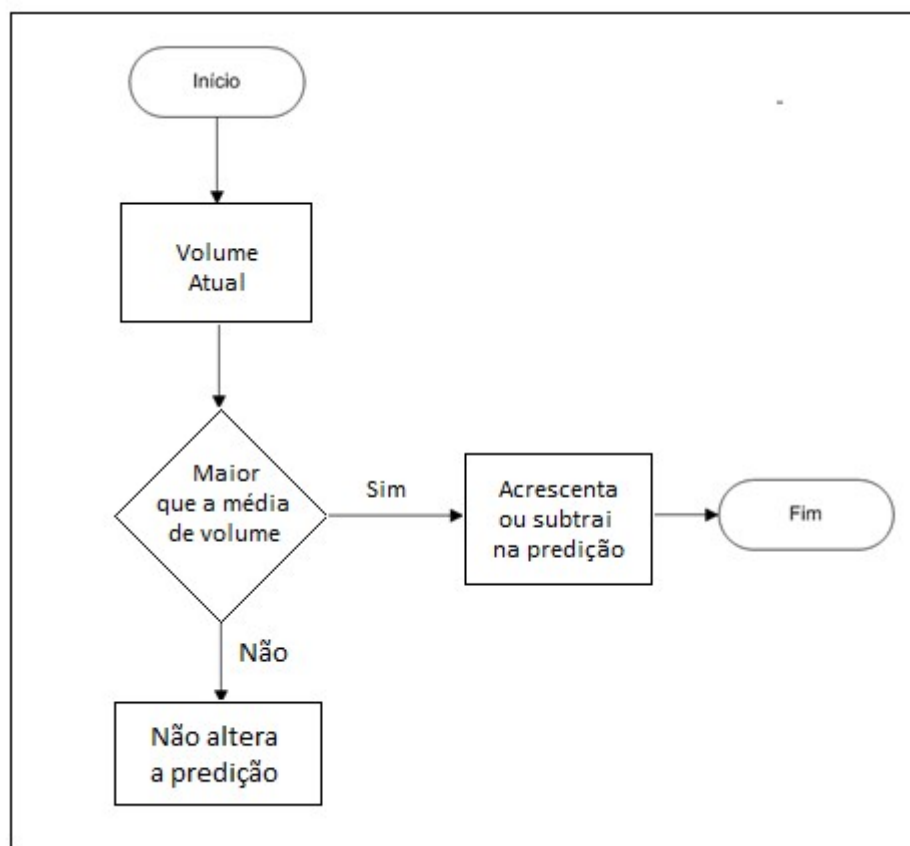
Pela Fórmula 4, o resultado do total de palavras coletadas dos textos, dividido pelo total de palavras e multiplicado pela quantidade de documentos dependerá do resultado da classificação dos textos. Se a classificação for positiva, será somado a predição já realizada, caso contrário, se a classificação for negativa, será subtraído.

No caso dos sites sofrerem atrasos nas atualizações das notícias, ou mesmo não serem atualizados, para não comprometer o resultado, a classificação pode permanecer, por exemplo, como positiva de forma estática, uma vez que o resultado da predição é associado à média de volume das ações negociadas.

Assim, como os preços das ações estão sendo coletados a cada minuto, também está sendo coletado o volume das ações negociadas, em quantidade de milhões. O volume das ações negociadas pode ter relação com as oscilações de preços que a ação sofre ao longo do dia.

Para associar o volume das ações às oscilações de preço, será calculada a média do volume das ações com o volume atual. O fluxograma da Figura 14 apresenta o processo decisório da média de volume das ações negociadas.

Fig. 10 - Fluxograma da média de volume das ações negociadas.



Fonte: Criado pelo autor.

O período de análise do volume foi de 30.01.2017 a 29.03.2017.

Se o volume atual for maior ou igual a média do volume analisado no período, o resultado da fórmula da aferição da Figura 14 será adicionado à predição, caso contrário, se o volume atual for menor que a média do volume das ações negociadas a predição não será alterada.

3.6 Coleta dos preços das ações

Os preços das ações variam a todo momento. Em determinado horário do dia, durante o pregão eletrônico, podem sofrer fortes ou fracas variações para cima ou para baixo, ou mesmo permanecer neutros. Seus movimentos são inconstantes.

A análise das oscilações dos preços é importante também para construção de histórico. Para este estudo, foram coletados os seguintes itens:

- Preço das ações da Ambev (Abev3);

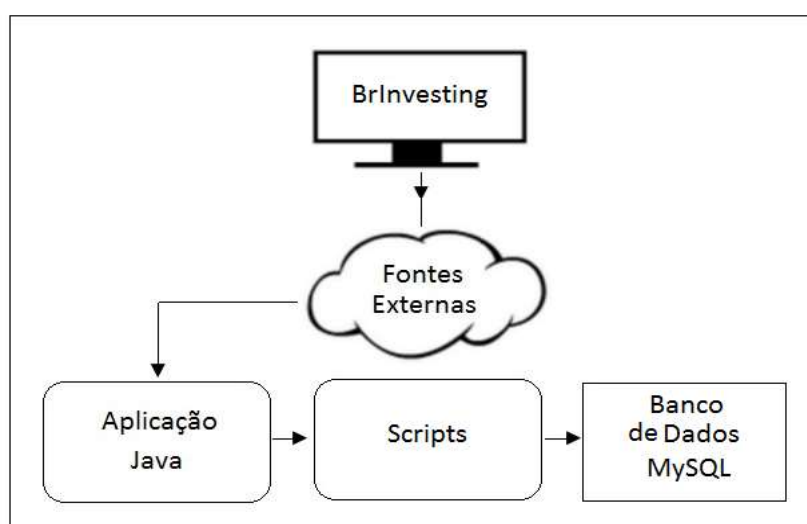
- Volume das ações da Ambev (Abev3);

Esses dois valores foram coletados a partir de 03.05.2016 a 01.07.2017. Esses valores são coletados pelo site <http://br.investing.com/indices/major-indices> de segunda à sexta-feira, das 10h às 17h, em intervalos de um em um minuto.

3.7 Programa coletor dos preços das ações da Ambev e volume das ações negociadas

O programa coletor de preços e volume negociado das ações da Ambev é similar ao programa de coleta de sites e foi desenvolvido em *Java*. Ele realiza leitura dos sites, sem a necessidade de abri-los como página *Web*. Os dados são convertidos como *HTML* e depois é gerado um arquivo texto.

Fig. 11 - Coleta de preços e volume da ação Abev3 e Ibovespa do site BrInvesting para composição de aferição de predição dos preços da ação da Ambev.



Fonte: Criado pelo autor

A figura 11 apresenta o programa coletor dos preços das ações da Ambev e volume das ações negociadas.

O arquivo texto gerado pelo programa coletor contém muitas informações desnecessárias, como detalhes sobre a estrutura da página *HTML*, cabeçalhos e muitas outras desnecessárias. Com a utilização de *scripts* automatizados, somente os preços das ações da Ambev e do Ibovespa são extraídos e registrados em um banco de dados relacional *MySQL*.

3.8 Análise de linhas

Quando o pregão eletrônico é iniciado, logo nos minutos iniciais, ocorre uma acomodação de preço referente ao fechamento do dia anterior. Por exemplo, quando do fechamento do pregão a ação Abev3 fechou em R\$ 19.40. No dia seguinte, normalmente e não de forma obrigatória, o pregão é aberto com o mesmo valor de R\$ 19.40. A partir do momento em que se iniciam as compras e vendas das ações, o preço começa a ser alterado, por volta das 10h13, aumentando ou diminuindo seu valor.

Tabela 5 - Dados do preço da ação da Ambev transportados para uma planilha para ilustrar a acomodação de preço do dia anterior.

		16-05	17-05	18-05	19-05	20-05	23-05	24-05	25-05	27-05	30-05	31-05	01-06	02-06	03-06	06-06
3																
7	10:03	18.76	18.77	18.56	18.40	18.77	18.54	18.44	18.69	18.76	19.30	19.15	19.11	19.34	19.28	19.50
8	10:04	18.76	18.77	18.56	18.40	18.77	18.54	18.44	18.69	18.76	19.30	19.15	19.11	19.34	19.28	19.50
9	10:05	18.76	18.77	18.56	18.40	18.77	18.54	18.44	18.69	18.76	19.30	19.15	19.11	19.34	19.28	19.50
10	10:06	18.76	18.77	18.56	18.40	18.77	18.54	18.44	18.69	18.76	19.30	19.15	19.11	19.34	19.28	19.50
11	10:07	18.76	18.77	18.56	18.40	18.77	18.54	18.44	18.69	18.76	19.30	19.15	19.11	19.34	19.28	19.50
12	10:08	18.76	18.77	18.56	18.40	18.77	18.54	18.44	18.69	18.76	19.30	19.15	19.11	19.34	19.28	19.50
13	10:09	18.76	18.77	18.56	18.40	18.77	18.54	18.44	18.69	18.76	19.30	19.15	19.11	19.34	19.28	19.50
14	10:10	18.76	18.77	18.56	18.40	18.77	18.54	18.44	18.69	18.76	19.30	19.15	19.11	19.34	19.28	19.50
15	10:11	18.76	18.77	18.56	18.40	18.77	18.54	18.50	18.75	18.76	19.30	19.15	19.01	19.34	19.28	19.50
16	10:12	18.62	18.77	18.56	18.40	18.77	18.54	18.50	18.75	18.76	19.21	19.15	19.05	19.30	19.28	19.50
17	10:13	18.67	18.73	18.49	18.30	18.73	18.33	18.51	18.73	18.76	19.21	19.11	19.04	19.29	19.33	19.50
18	10:14	18.71	18.77	18.48	18.33	18.77	18.43	18.50	18.72	18.76	19.19	19.06	19.03	19.23	19.31	19.50
19	10:15	18.71	18.74	18.51	18.38	18.74	18.36	18.52	18.74	18.76	19.22	19.06	19.04	19.23	19.32	19.50
20	10:16	18.77	18.75	18.55	18.37	18.75	18.35	18.53	18.69	18.76	19.23	19.05	19.05	19.21	19.32	19.50
21	10:17	18.77	18.75	18.49	18.38	18.75	18.41	18.51	18.71	18.76	19.23	19.03	19.05	19.14	19.32	19.50
22	10:18	18.74	18.72	18.50	18.32	18.72	18.41	18.48	18.70	18.76	19.26	19.04	19.06	19.14	19.35	19.50
23	10:19	18.76	18.74	18.49	18.30	18.74	18.40	18.46	18.72	18.76	19.25	19.02	19.08	19.11	19.33	19.52
24	10:20	18.76	18.74	18.48	18.37	18.74	18.36	18.45	18.69	18.76	19.25	19.04	19.06	19.12	19.33	19.60

Fonte: Criado pelo autor.

A Tabela 5, mostra o preço da ação da Abev3 da data 16.05.2016 a 06.06.2016. Apenas para ilustrar, os dados que encontram-se em Banco de Dados foram copiados e transportados para uma planilha. A linha em destaque, mostra que por volta das 10h13, em alguns dias desse intervalo, os preços começaram a se alterar..

Com o programa coletor do preço da ação da Abev3 sendo capturado de minuto a minuto, torna-se mais simples a identificação das variações do preço ao longo do dia, não é eficiente capturar os valores com intervalos maiores, de quinze em quinze minutos, por exemplo.

Os valores históricos dos preços foram utilizados na horizontal e não na vertical, assim a predição ocorrerá minuto a minuto e com vários dias diferentes e não por uma média diária por coluna. Isso foi possível devido à utilização da função *t()* *transpose* da linguagem e ambiente de estatísticas *R*. Para ilustrar a utilização da

função *transpose* foram copiados os horários 10h13min e 10h14min, dos dias 16.05.2016 a 02.06.2016, do banco de dados para uma planilha.

Fig. 12 - Utilização da função *transpose* da linguagem e ambiente de estatísticas *R* transpondo a matriz de preço da ação Abev3 para *data frame*.

10:13	18.67	18.73	18.49	18.30	18.73	18.33	18.51	18.73	18.76	19.21	19.11	19.04	19.29
10:14	18.71	18.77	18.48	18.33	18.77	18.43	18.50	18.72	18.76	19.19	19.06	19.03	19.23
...													
<div> <div> <div>...</div> <div>10:13</div> <div>10:14</div> <div>...</div> </div> <div> <div>18.67</div> <div>18.71</div> <div>18.73</div> <div>18.77</div> <div>18.49</div> <div>18.48</div> <div>18.30</div> <div>18.33</div> <div>18.73</div> <div>18.77</div> <div>18.33</div> <div>18.43</div> <div>18.51</div> <div>18.50</div> <div>18.73</div> <div>18.72</div> <div>18.76</div> <div>18.76</div> <div>19.21</div> <div>19.19</div> <div>19.11</div> <div>19.06</div> <div>19.04</div> <div>19.03</div> <div>...</div> <div>...</div> </div> </div>													

Função *t()* para transpor a matriz de valores da ação abev3 de linhas para colunas para data frame no R-Cran

Fonte: Criado pelo autor.

A função *t()* *transpose* transporá todas as linhas para colunas, aplicando essa ação às 420 linhas que representam a quantidade de minutos para o horário do pregão eletrônico, das 10h às 17h. No banco de dados não existem várias colunas com os minutos, mas apenas uma única coluna chamada “último” que representa a cotação do último preço da ação Abev3, atualizada minuto a minuto. O programa em *R* realiza a pesquisa *SQL* com os valores dos preços, transpondo-os para utilização da predição. O resultado da função *t()* *transpose* é utilizado em *Data Frames*.

3.9 Predição do valor da ação Abev3 com classificador *Naive Bayes*

O classificador *Naive Bayes* é baseado no Teorema de *Bayes*. Com o teorema de *Bayes* é possível calcular a probabilidade posterior, $P(c|x)$, de $P(c)$, $P(x)$, e $P(x|c)$. *Naive Bayes* utiliza o aprendizado *bayesiano* para treinar os classificadores. O aprendizado *Bayesino* utiliza um modelo probabilístico baseado no conhecimento prévio do problema, combinado com exemplos de treinamento, para determinar a probabilidade final de uma hipótese.

Fórmula 5 - Classificador Naive Bayes com variáveis C e X.

$$P(c|x) = \frac{P(x|c) P(c)}{P(x)}$$

Fonte: (SANTIDHANYAROJ, 2015) Adaptado e traduzido para o português pelo autor.

Onde:

$P(c|x)$: probabilidade da classe c dado o vetor x

$P(x|c)$: probabilidade do vetor x dado a classe c

$P(c)$: probabilidade *a priori* da classe c

$P(x)$: probabilidade *a priori* do vetor de treinamento x

Para a realização do cálculo da probabilidade $P(c|x)$, é necessária a construção das tabelas de frequência e de probabilidade e, também, dos valores das ações que foram coletados na seção 3.8. Na Tabela 6, por exemplo, foram utilizados os valores da ação Abev3 de 13 dias de segunda à sexta-feira nas datas de 03.05.2016 a 19.05.2016 no período das 11h. O sim na última linha em vermelho é o resultado da predição.

A coluna Quantidade, apenas indica a quantidade de dias analisados nesse exemplo, no caso 13 dias. A coluna Data, apresenta os dias úteis da semana que foram analisados, e a coluna Abev3 apresenta os valores de cada um dos 13 dias, em que é mantido o valor de R\$ 19.43, inicialmente, depois uma acentuada queda chegando a R\$ 18.77, então uma subida até R\$ 19.05, e, na sequência, uma forte queda e atingindo R\$ 18.59 e baixando até R\$ 18.51. Na figura 15, é apresentada, graficamente, a variação durante esse período.

O problema atribuído ao *Naive Bayes* como pergunta principal à predição do valor da ação Abev3 é: O preço da ação Abev3 para o décimo quarto dia, subirá no período de análise?

A última coluna chamada Subirá, indica a condição que se encontra a atual cotação do valor da ação Abev3, sendo estabelecidos os seguintes critérios:

- valor da cotação atual é igual ao valor anterior, a coluna foi preenchida com o resultado “Manteve”;
- valor da cotação atual é maior que valor anterior, a coluna foi preenchida com o resultado “Sim”;
- valor da cotação atual é menor que valor anterior, a coluna foi preenchida com o resultado “Não”.

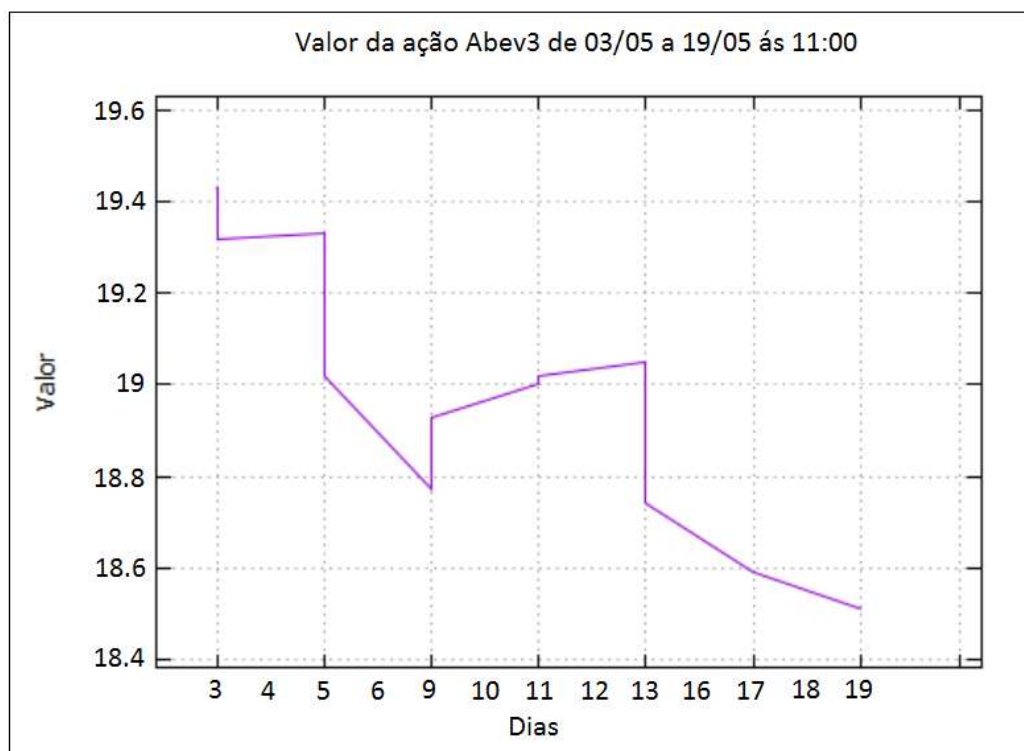
Tabela 6 - Tabela com os valores da ação da Abev3 de 03 a 19/05 de 2016 coletados às 11h.

Quantidade	Data	Abev3	Subirá
1	03/05/2016	19.43	Manteve
2	04/05/2016	19.32	Não
3	05/05/2016	19.33	Sim
4	06/05/2016	19.02	Não
5	09/05/2016	18.77	Não
6	10/05/2016	18.93	Sim
7	11/05/2016	19.00	Sim
8	12/05/2016	19.02	Sim
9	13/05/2016	19.05	Sim
10	16/05/2016	18.74	Não
11	17/05/2016	18.59	Não
12	18/05/2016	18.59	Manteve
13	19/05/2016	18.51	Não
14	20/05/2016	18.59	Sim

Fonte: Criado pelo autor.

Os dados da Tabela 6 foram extraídos apenas para ilustração, via *SQL* (*Structure Query Language*) e, construída a tabela. Os dados são coletados em intervalos de 1 em 1 minuto, de forma automática, durante a predição com a utilização da linguagem e ambiente de estatísticas *R*.

Fig. 13 - Gráfico com os valores da ação da Abev3 de 03 a 19/05 de 2016.



Fonte: Criado pelo autor.

O gráfico da figura 13, apresenta uma forte queda no valor da ação Abev3 e somente nos dias entre 09 a 13 houve uma subida, mas voltando a cair entre os dias 13 a 19.

A tabela de frequência faz a contagem de todos os 'sim' e 'não' das possíveis variações dos preços das ação Abev3: sobe, desce e manteve.

Tabela 7 - Tabela de frequência para composição da predição do valor da ação da Abev3 de 03 a 19/05 de 2016.

Abev3	Sim	Não
Sobe	5	
Desce		6
Manteve	2	
Total	7	6

Fonte: Criado pelo autor.

O Total das respostas 'sim' e 'não' será utilizado na fórmula *Naive Bayes* para o cálculo do $P(\text{Sim}|\text{Subirá})$.

A Tabela 8 apresenta as probabilidades para cada variação do preço da ação Abev3 em aumentar, diminuir ou manter-se o mesmo do exemplo estudado. Analisando isoladamente somente a variação de subir, o percentual é de 0.38 ou 38% de chances do preço da ação subir, sendo inferior ao de descer que corresponde a 46% e 15% de se manter o mesmo valor.

Tabela 8 - Tabela de probabilidade para composição da predição do valor da ação da Abev3 de 03 a 19/05 de 2016.

Abev3	Sim	Não		
Sobe	5		5/13	0.38
Desce		6	6/13	0.46
Manteve	2		2/13	0.15
Total	7	6		
	7/13	6/13		
	0.53	0.46		

Fonte: Criado pelo autor.

Considerando somente a opção 'sim' o resultado é 53% contra 46% de 'não' ou de 'descer', indicando que o preço da ação subirá..

Para se estimar a predição é necessário calcular o $P(\text{Sim}|\text{Subirá})$ pela fórmula do classificador *Naive Bayes*.

Até a medição 13, que contou com 13 análises de preços da ação Abev servindo de histórico e parâmetros para a fórmula de *Naive Bayes* ser calculada, não era possível identificar o que ocorreria na medição 14, se a ação subiria, desceria ou se manteria. Substituindo os valores extraídos das tabelas de frequência e probabilidade e aplicado a fórmula de *Naive Bayes*, foi possível identificar um percentual de 97% de chances do valor da ação Abev3 subir. Essa predição se confirmou na Tabela 6, última linha em vermelho, subindo o preço da ação de R\$ 18.51 para R\$ 18.59.

Fig. 14 - Exemplo de cálculo da probabilidade do preço da ação Abev3 subir no período analisado 03 a 19/05 de 2016.

$$P(\text{Sim}|\text{Subirá}) = \frac{P(\text{Subirá}|\text{Sim}) * P(\text{Sim})}{P(\text{Subirá})}$$

$$P(\text{Subirá}|\text{Sim}) = 5/7$$

$$P(\text{Sim}) = 7/13$$

$$P(\text{Subirá}) = 5/13$$

Substituindo:

$$P(\text{Sim}|\text{Subirá}) = \frac{5/7 * 7/13}{5/13} \Rightarrow \frac{0,71 * 0,53}{0,38} = 0,97$$

Fonte: Criado pelo autor.

Mesmo com um resultado de 97% de probabilidade de uma predição acontecer, neste caso informando que o preço da ação subiria, não existe certeza de que o preço da ação subirá de fato. Também não foram levados em consideração outros fatores que poderiam inverter o resultado reduzindo o preço. Esses fatores poderiam ser acontecimentos extraordinários como tragédias, acidentes, divulgação de balanço negativo, divulgação de escândalos envolvendo a companhia, mudanças governamentais, ameaças de guerras entre outros.

Outro fator que pode contribuir para uma predição incorreta é a quantidade de eventos analisados. No exemplo, foram analisados apenas 13 dias, o que pode ser insuficiente para uma predição com maior precisão.

Este exemplo apenas mede o horário das 11h em 13 dias para ilustração da aplicação da fórmula matemática. Para a predição completa, são usados os horários das 10h às 17h, de segunda à sexta-feira, com um histórico de 224 dias úteis.

No dia 09 de setembro de 2016, os mercados mundiais estavam em estado de cautela, pois os juros americanos poderiam subir. Além dessa notícia, a Coreia do Norte realizou seu quinto e mais potente teste nuclear, causando quedas nas bolsas do mundo todo. (OLIVEIRA, 2016)

Esse evento gerou uma anormalidade nos preços da ação da Abev3. A predição realizou os cálculos com histórico de vinte dias e no vigésimo primeiro dia, que seria o dia 09 de setembro, ocorreu o teste nuclear da Coreia do Norte que não foi adicionada à predição. Na Figura 17 é apresentado o cálculo pela fórmula de *Naive Bayes* às 10h37min, gerando uma predição de 98% do preço da ação Abev3 subir, fato que não aconteceu. No dia 09 de setembro, houve quedas nos valores da ação que permaneceram em baixa boa parte do dia.

Tabela 9 - Tabela com os valores da ação da Abev3 de 11.08.2016 a 09.09.2016 às 10h37min.

Quantidade	Data	Abev3	Subirá
1	11/08/2016	19.26	Manteve
2	12/08/2016	19.17	Não
3	15/08/2016	19.48	Sim
4	16/08/2016	19.69	Sim
5	17/08/2016	19.69	Manteve
6	18/08/2016	19.91	Sim
7	19/08/2016	19.87	Não
8	22/08/2016	19.76	Não
9	23/08/2016	19.78	Sim
10	24/08/2016	19.62	Não
11	25/08/2016	19.48	Não
12	26/08/2016	19.71	Sim
13	29/08/2016	19.35	Não
14	30/08/2016	19.32	Não
15	31/08/2016	19.19	Não
16	01/09/2016	19.18	Não
17	02/09/2016	19.05	Não
18	05/09/2016	19.80	Não
19	06/09/2016	19.85	Sim
20	08/09/2016	19.90	Sim
21	09/09/2016	19.89	Não

Fonte: Criado pelo autor.

Tabela 10 - Tabela de probabilidade para composição da predição do valor da ação da Abev3 de 11.08.2016 a 09.09.2016.

Tabela de Frequência			Tabela de Probabilidade			
Abev3	Sim	Não	Abev3	Sim	Não	
Sobe	7		Sobe	7		7/21 0.33
Desce		12	Desce		12	12/21 0.57
Manteve	2		Manteve	2		2/21 0.09
Total	9	12	Total	9	12	
				9/21	12/21	
				0.42	0.57	

Fonte: Criado pelo autor.

Fig. 15 - Cálculo da probabilidade do preço da ação Abev3 subir no período analisado 11.08.2016 a 09.09.2016.

$$P(\text{Sim} | \text{Subirá}) = \frac{P(\text{Subirá} | \text{Sim}) * P(\text{Sim})}{P(\text{Subirá})}$$

$$P(\text{Subirá} | \text{Sim}) = 7/9$$

$$P(\text{Sim}) = 9/21$$

$$P(\text{Subirá}) = 7/21$$

Substituindo:

$$P(\text{Sim} | \text{Subirá}) = \frac{7/9 * 9/21}{7/21} \Rightarrow \frac{0,77 * 0,42}{0,33} = 0,98$$

Fonte: Criado pelo autor.

Como *Big Data* é possível extrair os textos das notícias dos sites financeiros e com o processamento de linguagem natural identificar os “sentimentos” dos textos que seriam positivos, negativos ou neutros.

Eventos extraordinários como o da Coreia do Norte precisam estar registrados dentre as palavras-chaves dos dicionários positivos e negativos para serem identificados e calculados os seus respectivos resultados.

3.10 Predição com linguagem e ambiente de estatísticas R

A linguagem e ambiente de estatísticas *R* é um programa estatístico e gráfico. Com adições de bibliotecas externas é possível expandir sua capacidade para atender a algum propósito específico. A biblioteca externa que adiciona funções para realização de predições é chamada de pacote *E1071*, com ela é possível realizar as predições chamando a função *naiveBayes*, descrita na documentação dessa classe.

O processo de predição realizado com a linguagem e ambiente de estatísticas *R* utiliza todos os minutos desde a abertura até o fechamento do pregão, incluindo 224 dias de dados. Dessa forma, minuto a minuto, no período das 10h às 17h de segunda à sexta-feira, ou 420 minutos por dia são processados para a predição.

A predição não precisa ocorrer em tempo real, diferentemente das coletas do Big Data. Ela pode ser executada após o fechamento do pregão no dia anterior, com foco para o dia seguinte, ou até mesmo até minutos antes de se iniciar a abertura da bolsa. Para isso, é necessário possuir os valores dos preços da ação Abev3 do dia anterior, extraídos pelo programa coletor.

O resultado da predição possibilita ao investidor ter uma visão do que aconteceria ao longo do dia com antecedência. A predição gera, antecipadamente, um gráfico das variações minuto a minuto no horário do pregão, permitindo que se visualizem os valores máximo e mínimo que o preço da ação pode atingir, além de identificar um melhor momento para compra e venda, facilitando o planejamento do investidor.

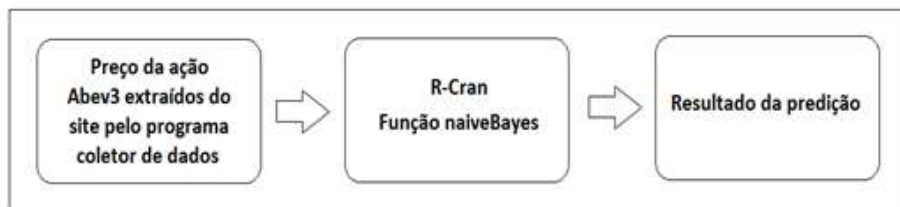
Com a predição calculada um dia antes, ou mesmo minutos antes da abertura da bolsa, o valor apresentado não reflete as variações e oscilações de mercado que o preço da ação pode sofrer durante o pregão eletrônico. Podemos citar todas as variações de preços em que as notícias e acontecimentos publicados na Internet afetaram o preço da ação

Um cenário de incertezas, que pode alterar a predição que foi realizada previamente, os cálculos efetuados pela linguagem e ambiente de estatísticas *R* ajustam o valor do preço da predição em tempo real, proporcionando o conhecimento do melhor valor possível do preço da ação para o investidor.

Na Figura 18 é apresentado o processo macro de predição realizado, pela linguagem e ambiente de estatísticas *R*. Para a realização desse processo, foi

desenvolvido um *script* com a linguagem e ambiente de estatísticas *R* de aproximadamente 12.000 linhas de código.

Fig. 16 - Processo macro de predição pela linguagem e ambiente de estatísticas *R*.



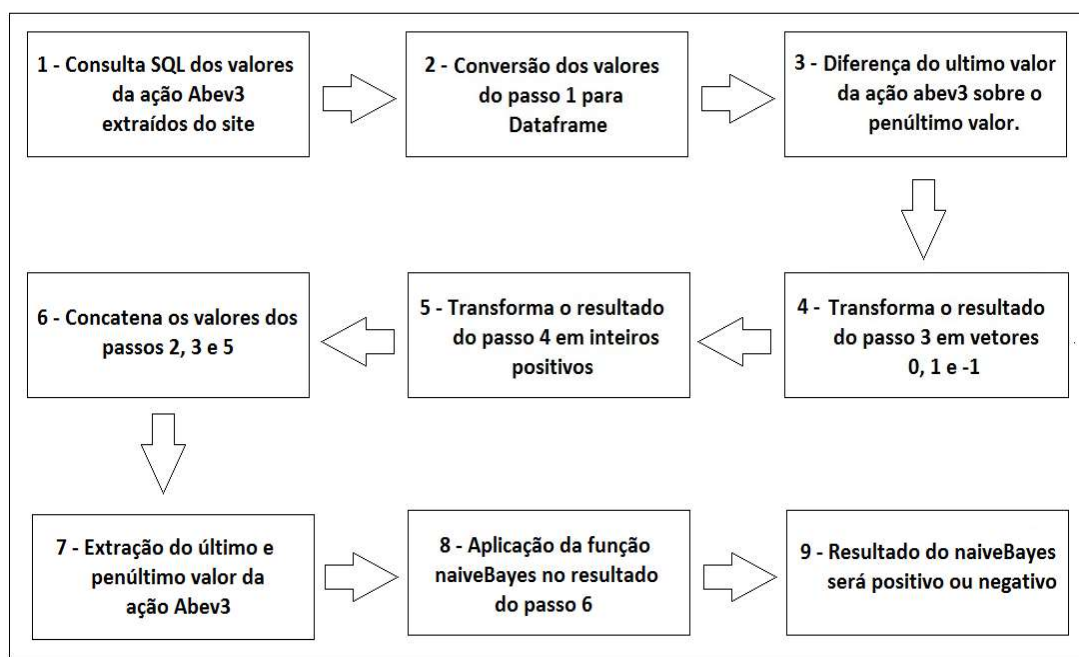
Fonte: Criado pelo autor.

O preço da ação da Abev3 é extraído pelo site <http://brinvesting.com> pelo programa coletor de dados com intervalo de 60 segundos entre as coletas.

3.11 Detalhamento da função *naiveBayes*

A Figura 19 apresenta o processo de predição detalhado. As etapas 1 a 9 são sequenciais e realizadas para cada valor do preço da ação que foi extraído minuto a minuto. Esse processo se repete para cada minuto ou 420 vezes para a realização da predição inteira.

Fig. 17 - Etapas para a realização da predição com a função *naiveBayes*.



Fonte: Criado pelo autor.

Etapa 1: Extração dos valores da ação Abev3 pelo site

Os valores dos preços das ações são extraídos do site <http://brinvesting.com> e armazenados em banco de dados, mas para a predição é utilizado o preço do dia anterior mais o valor histórico do dia que começou em 03.05.2016. Para a predição não é utilizado o preço em tempo real. A consulta do preço da ação é realizada via *SQL*, executado diretamente na linguagem e ambiente de estatísticas *R* com a utilização do pacote *RMySQL*.

Passo 2: Conversão dos valores do passo 1 para *Dataframe*

A linguagem e ambiente de estatísticas *R* não permite que se use um valor numérico sem que ele não esteja de acordo com as padrões numéricos que a ferramenta utiliza. Muitas vezes são necessárias conversões quando um valor é pesquisado de um banco de dados ou lidos por arquivos. Os valores do passo 1 foram convertidos para o formato *Dataframe*, possibilitando que os dados fossem armazenados na forma de tabela.

Passo 3 - Diferença do último valor da ação Abev3 sobre o penúltimo valor

Subtraindo-se o último valor da ação Abev3 sobre o penúltimo, encontra-se a variação do valor positivo ou negativo. Se positivo, o preço da ação subiu, se negativo, o preço da ação desceu.

Passo 4 - transforma o resultado do passo 3 em vetores, 0, 1 e -1. O resultado do passo 3 é convertido em vetores com a função *sign()* obtendo-se, assim, os valores 1 e -1 para os valores positivos, negativos e repetidos. Tal conversão é necessária para passo 5.

Passo 5 - transforma o resultado do passo 4 em inteiros positivos. Retorna os inteiros positivos, testando e retornando a condição *TRUE* ou *FALSE*.

Passo 6 - concatena os valores dos passos 2, 3 e 5. Os resultados dos passos 2, 3 e 5 são concatenados em *Dataframe* com a função *cbind*.

Passo 7 - extração do último e penúltimo valor do passo 6. Este passo separa o último e o penúltimo valores para o cálculo da predição. O resultado desse passo será utilizado para o cálculo da predição.

Passo 8 – aplicação da função *naiveBayes*. Nesse passo será aplicada a função *naiveBayes* da tabela concatenada do passo 6.

Fig. 18 - Fragmento de código para processamento da predição com a função *naiveBayes* pela linguagem e ambiente de estatísticas R e pacote E1071.

```
> col321.f10 <- naiveBayes(col321.f5 ~ ., col321.f6)
> col321.f11 <- predict(col321.f10, col321.f6$col321.f5, type = "raw")
```

Fonte: Criado pelo autor.

A função *naiveBayes* é aplicada nos resultados do passo 5 e concatenada no passo 6. Na figura 20, é apresentado um fragmento de codificação que mostra a linha *col321.f10* recebendo a função *naiveBayes*. Logo abaixo, a linha *col321.f11* recebe o resultado da função *naiveBayes* e da linha *col321\$col321.f5*, representando a linha para inteiros positivos, calculando, assim, a predição.

Passo 9 – Resultado do *naiveBayes*

Ao fim do cálculo da predição, será apenas informado se a predição foi positiva ou negativa. Na figura 21 é apresentado um resultado de exemplo de uma predição sendo 0,61 ou 61% de chances do valor da ação subir e 0,38 % de não subir.

Fig. 19 - Apresentação da função *naiveBayes* pela linguagem e ambiente de estatísticas *R* e pacote *E1071*.

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
      FALSE      TRUE
0.6105263 0.3894737
```

Fonte: Criado pelo autor.

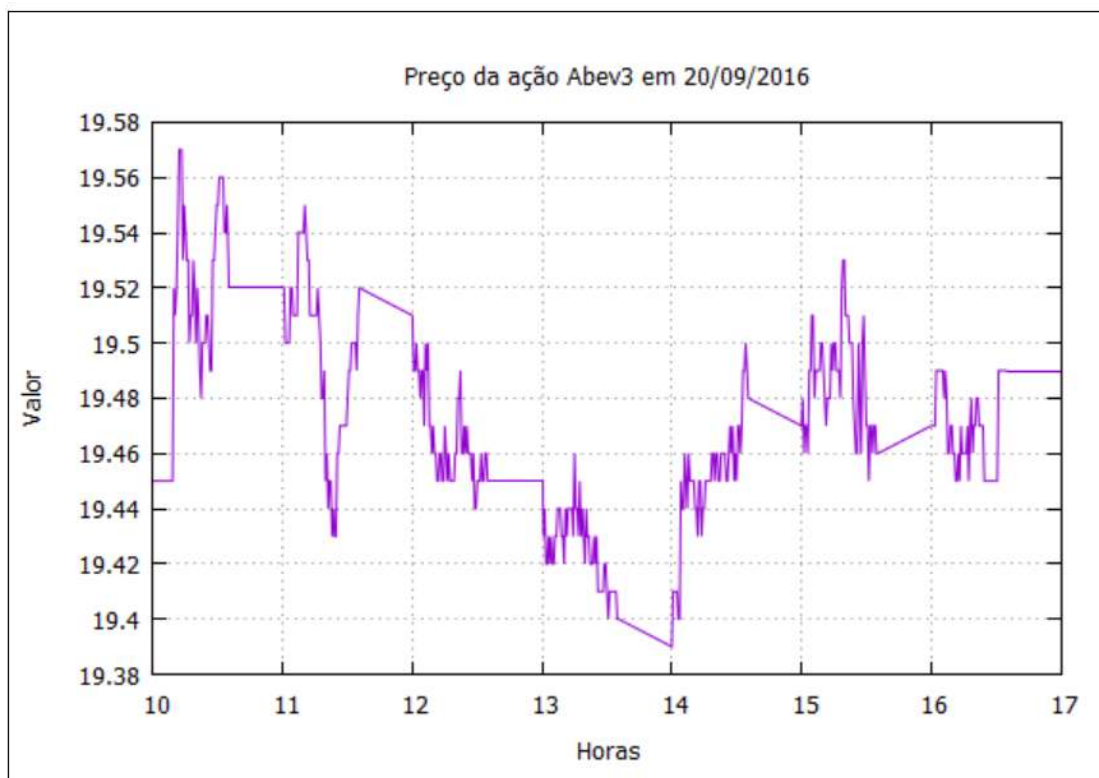
Este resultado não é apresentado para cada cálculo realizado pelo passo 9. O resultado da função *naiveBayes* apenas retorna positivo ou negativo.

3.12 Cálculo de predição

Segundo (LEMOS, 2015) os preços das ações seguem movimentos aleatórios impossibilitando previsões precisas. Talvez não exista uma formula matemática capaz de prever os preços com absoluta precisão e certeza.

Na Figura 22 é apresentado um gráfico com as variações dos preços da ação Abev3 das 10h às 17h. Houve muitas variações ao longo do dia, sendo que a máxima ultrapassou o valor de R\$ 19.56 e a mínima R\$ 19.40.

Fig. 20 - Preço da ação Abev3 em 20/09/2016 das 10 às 17:00.



Fonte: Criado pelo autor.

O gráfico mostra as variações dos preços minuto a minuto. Neste gráfico, não houve eventos extraordinários, que provocasse mudanças bruscas para cima ou para baixo no valor da ação, em um prazo de tempo bastante curto. O gráfico mostra as variações minuto a minuto em um dia regular no pregão.

Neste dia, as variações dos valores não ultrapassaram a R\$ 0,04 para cima e para baixo. A cada minuto, os preços variavam de R\$ 0,01 a R\$ 0,03 e R\$ -0,01 e R\$ -0,03 na maior parte do dia. Observando que essas variações, não ultrapassaram a diferença do último valor (cotação mais atual) e da penúltima cotação.

Tabela 11 - Diferença do último e penúltimo valor do 12:05 às 12:02.

Horário	Valor Atual	Diferença último valor pelo penúltimo
12.05	19,49	0,01
12.06	19,48	-0,01
12.07	19,49	0,01
12.08	19,49	0,00
12.09	19,47	-0,02
12.10	19,50	0,03
12.11	19,49	-0,01
12.12	19,50	0,01

Fonte: Criado pelo autor.

A diferença da cotação mais atual para o penúltimo valor acompanha o preço da ação ao longo do dia. Por exemplo, na parte mais alta do gráfico que ocorreu às 10h20 e atingiu a cotação de R\$19.57 e seu anterior, às 19h19, atingiu R\$ 19.54, variando R\$ 0,03. O mesmo aconteceu na parte mais baixa do gráfico, às 14h e atingiu o valor de R\$ 19.39 e seu anterior às 13h59 estava em R\$ 19.40 variando R\$ 0,01.

A variação do último valor da ação para o penúltimo valor mostra uma relação entre os valores que serão utilizado na fórmula de predição.

O resultado da função *naiveBayes* ao final dos cálculos, informa se a predição foi *TRUE* ou *FALSE*, ou verdadeiro ou falso. A partir do resultado de predição, é necessária a realização do cálculo para cada uma das possibilidades da predição resultar em negativa. Visando à definição de medidas diferentes para os resultados positivos ou negativo

Abaixo é apresentada a fórmula de predição.

Fórmula 6 - Cálculo de predição.

$$\begin{array}{c}
 \text{Cálculo de} \\
 \text{Predição}
 \end{array}
 =
 \begin{array}{c}
 \text{Resultado da} \\
 \text{Predição}
 \end{array}
 \left[\begin{array}{l}
 \text{Valor atual} + (\text{Valor atual} - \text{Penúltimo valor}) \\
 \text{ou} \\
 \text{Valor atual} - (\text{Valor atual} - \text{Penúltimo valor})
 \end{array} \right]
 =
 \begin{array}{c}
 \text{Resultado do} \\
 \text{Cálculo de Predição}
 \end{array}$$

Fonte: Criado pelo autor

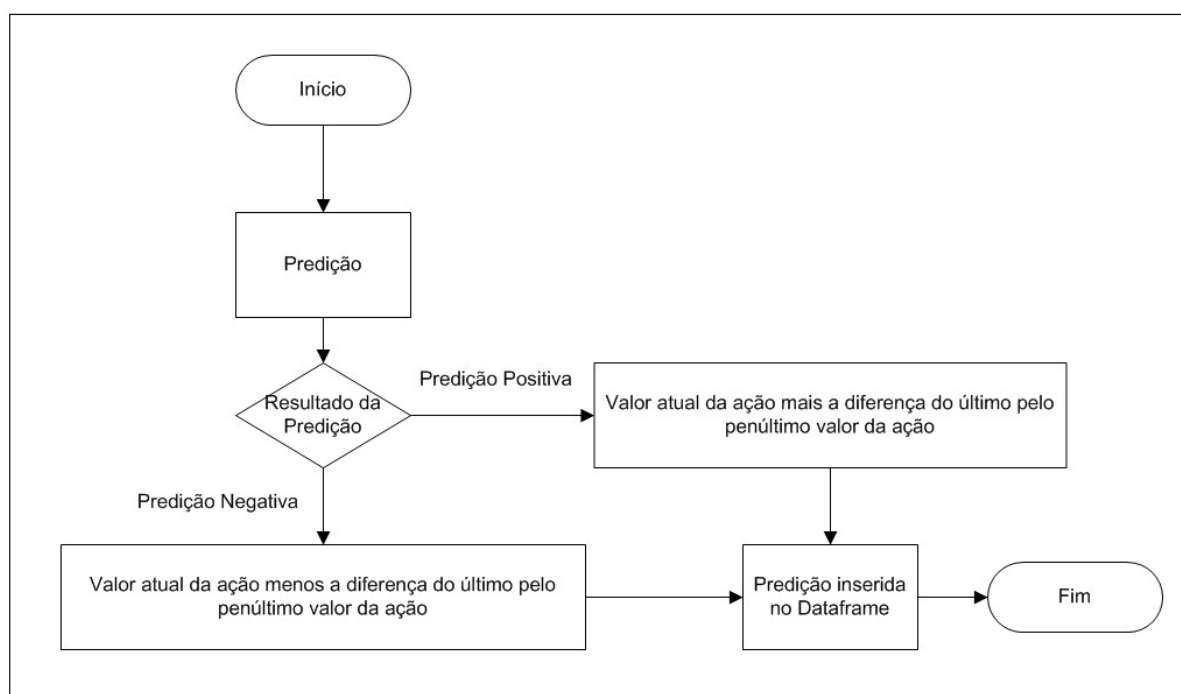
Após a função *naiveBayes* informar o resultado, sendo positivo ou negativo, a fórmula da predição conseguirá calcular os valores. Se o resultado for positivo, será

subtraído o valor atual do penúltimo valor e adicionado ao valor atual. Se o resultado for negativo, será subtraído o valor atual do penúltimo valor e subtraído do valor atual.

Essa fórmula está inserida em um *script* na linguagem e ambiente de estatísticas *R*. Para cada minuto durante o pregão eletrônico, essa fórmula é executada, gerando a predição.

A Figura 21 apresenta o fluxograma para realização do processo de predição. Para cada resultado da predição, ocorrerá um cálculo distinto.

Fig. 21 - Fluxograma para o cálculo de predição.



Fonte: Criado pelo autor.

O resultado da predição gerará um *Dataframe* de 420 linhas com os respectivos valores do preço da ação Abev3. Com esses valores, será possível construir um gráfico de predição. Este gráfico de predição representa a predição dos valores que a ação Abev3 do dia seguinte. A partir desse gráfico o investidor poderá visualizar os valores máximo e mínimo que o preço da ação poderá atingir dentro da predição e também o melhor momento para compra e venda.

4 AVALIAÇÃO DOS RESULTADOS

Neste capítulo são apresentados os resultados da pesquisa deste trabalho. Ele será dividido em três partes, sendo o resultado da predição com *Naive Bayes*, o resultado do *Big Data* com o processamento de linguagem natural e os dois resultados.

4.1 Resultado da predição *Naive Bayes*

Para a realização dessa parte do experimento foi analisado o período de 03.05.2016 a 29.03.2017 em que foram armazenados os valores das ações da empresa Ambev, a cada minuto, de segunda à sexta das 10h às 17h.

Com um total de 188 608 registros, na Tabela 11 são apresentados o histórico do valor da ação ABEV3, o valor da predição, a diferença da ação em relação à predição e a média da predição.

A média da predição é a soma de todos os valores coletados para análise e dividida pela quantidade de linhas. Como resultado, quanto mais próximo de 0, mais próximo estará do valor real do preço da ação.

Tabela 12 - Análise do histórico dos valores dos preços da ação e da predição.

Dias	Valor da ação	Valor da predição	V. ação - V. predição	Média da predição
1	19,49	19,57	-0,08	
2	19,25	19,44	-0,19	
3	18,94	19,12	-0,17	
4	18,86	18,79	0,07	
~	~	~	~	~
73	19,84	19,51	0,33	Média 0,09 - Bloco 1
74	19,84	19,76	0,07	
75	19,96	19,84	0,13	
76	19,89	19,91	-0,02	
~	~	~	~	~
148	16,07	16,57	0,50	Média 0,02 - Bloco 2
149	16,52	16,57	-0,05	
150	16,70	16,42	0,29	
151	16,62	16,62	0,00	
~	~	~	~	~
222	18,05	17,80	0,25	Média 0,12 - Bloco 3
223	18,20	17,98	0,23	
224	18,27	18,14	0,13	
Média	18,34	18,27	0,07	

Fonte: Autor

Comparando-se a média do valor da ação Abev3 em 224 dias medidos obtém-se R\$ 18,34 e a média da predição R\$ 18,27 ou - 0,38% abaixo do valor real.

A média foi calculada pela soma dos valores diários alcançados pela ação Abev3 divididos por 420 sendo este a quantidade de ocorrências durante o período do pregão eletrônico das 10:00 às 17:00 horas. Com esse valor para cada dia, foram somados e divididos por 224 dias obtendo-se a média geral. Assim, a média geral da diferença entre o valor real e a predição alcançada foi de -0,07 centavos.

Para identificar em que momento a predição apresentou o melhor resultado, foram divididas as 224 medições em dois blocos de 75 registros e um bloco de 74 registros. A média da diferença da predição e o valor da ação é:

- Primeiro bloco: 0,09
- Segundo bloco: 0,02
- Terceiro bloco: 0,12

O segundo bloco, com medição de 76 a 149 registros foi o que apresentou a predição mais próxima do valor real, seguido pelo primeiro e pelo último bloco.

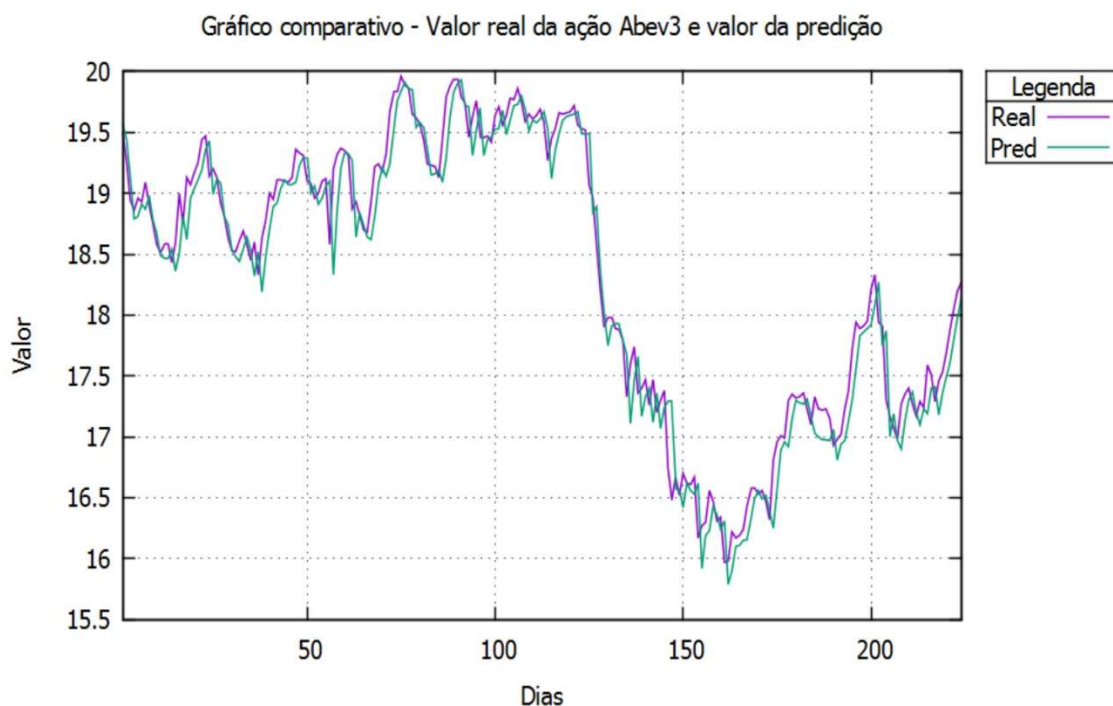
Uma das hipóteses para segundo bloco apresentar o melhor resultado foi o cenário político-econômico que o Brasil enfrentou em 2016-2017 com o impeachment da presidente Dilma e outros fatores políticos que afetaram diretamente o mercado.

A partir da análise desses números, um histórico maior não necessariamente resultaria em melhores predições. A análise em que apenas se verifica o resultado da predição com *Naïve Bayes*, não leva em considerações eventos extraordinários que podem afetar a normalidade do histórico dos preços da ação, como, por exemplo, o os cenários político e econômico no Brasil, em 2016 houve o *impeachment* da presidente, assim, o histórico se mostrou instável.

O gráfico da figura 26 apresenta a média dos 224 dias de medição do preço da ação Abev3 e sua predição.

Aa linha vermelha representa o preço da ação e a linha verde a predição alcançada.

Fig. 22 - Gráfico comparativo valor real da ação Abev3 e valor da predição.



Fonte: Criado pelo autor.

Com a predição sendo executada até um dia antes do pregão o resultado da predição alcançou valores próximos ao valor da ação, ficando em média R\$ 0,07 centavos abaixo.

- Total de registros: 188 608.
- Total de dias úteis medidos: 224.
- Média de acertos da predição considerando até 1 % em relação ao valor real da ação: 68,17%.
- Média de acertos da predição acima de 1 % em relação ao valor real da ação: 31,83 %.
- Média da predição em relação ao valor real da ação: manteve-se 0,07 centavos abaixo do valor real.
- Variação máxima no valor real: R\$ 15,80 a R\$ 20,07.
- Variação máxima na predição: R\$ 15,55 a R\$ 20,25.
- Valor da predição igual ao valor real: 1600 acertos ou 1.69 %.

4.2 Resultado do *Big Data* e processamento de linguagem natural

Para a realização do experimento do *Big Data* e processamento de linguagem natural foi analisado o período de 30.01.2017 a 29.03.2017. Nesse período foram extraídos 21778 textos totalizando 2.9 *Gbytes* de informações dos seis sites financeiros em intervalos de 15 minutos.

Como resultado a classificação de textos atingiu um percentual de 48.34 % ou 10.527 de 21.778 textos classificados corretamente. A verificação para a classificação de textos ocorreu de forma manual.

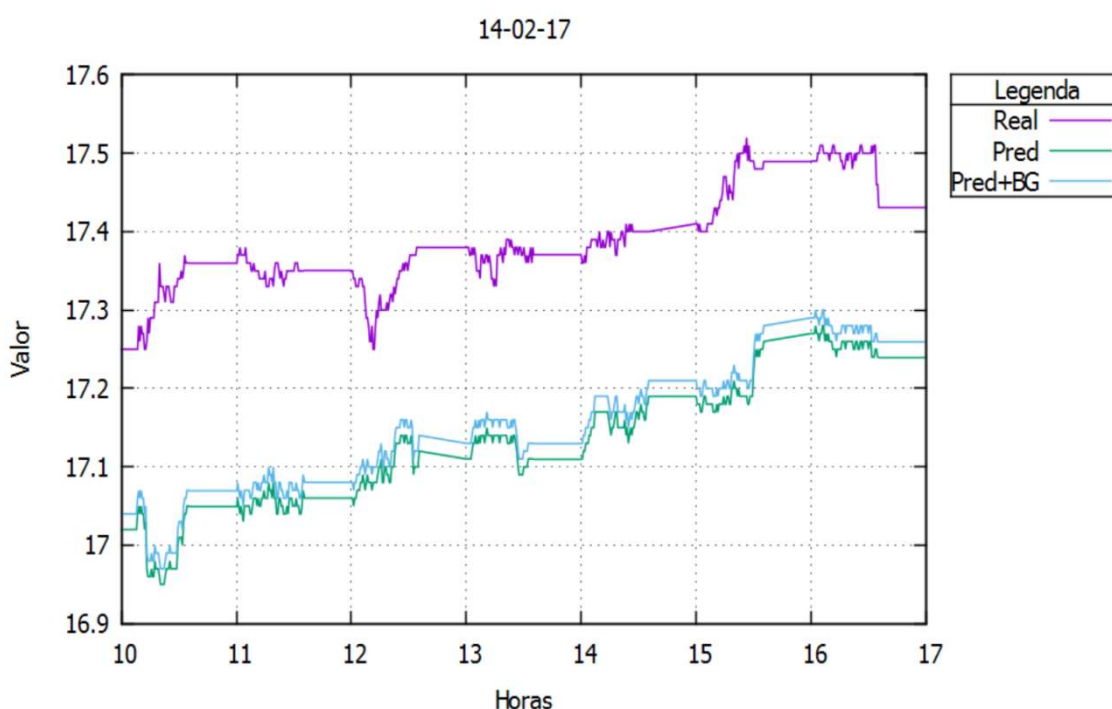
A fórmula de aferição de preços apresentou como resultado uma fração de centavos, que corresponde a valores muito pequenos e seu resultado pouco influenciou a predição. Como alternativa de correção, o resultado da fórmula passou

a ser multiplicado por 10, de forma automática, apresentando resultados mais próximos da realidade.

No geral, a maioria dos textos resultou em “sentimento positivo” e acrescentou em média R\$ 0,02 centavos à predição, aproximando-se do valor real.

No gráfico da Figura 27, é apresentada a oscilação de preço da ação Abev3 durante o pregão eletrônico do dia 14 de fevereiro de 2017.

Fig. 23 - Gráfico do dia 14/02/2017 apresentando o ajuste de R\$ 0,02 à predição aproximando-se do valor da ação Abev3.



Fonte: Criado pelo autor

A linha vermelha representa o valor real da ação Abev3. A linha verde, representa a predição calculada previamente. A linha azul, representa a predição mais o resultado da classificação de texto. Nesse caso, sendo positivo e sofrendo um ajuste de R\$0,02 pela fórmula de aferição, acrescentado ao valor da predição. Com esse acréscimo de R\$0,02 a predição apresentou um resultado mais próximo do valor real, se comparado somente à predição, sem a influência da classificação de textos.

Em um outro exemplo no dia 10 de fevereiro de 2017, a predição cruzou a linha, indicando um valor mais alto que o valor da ação. Por volta das 14h o preço da

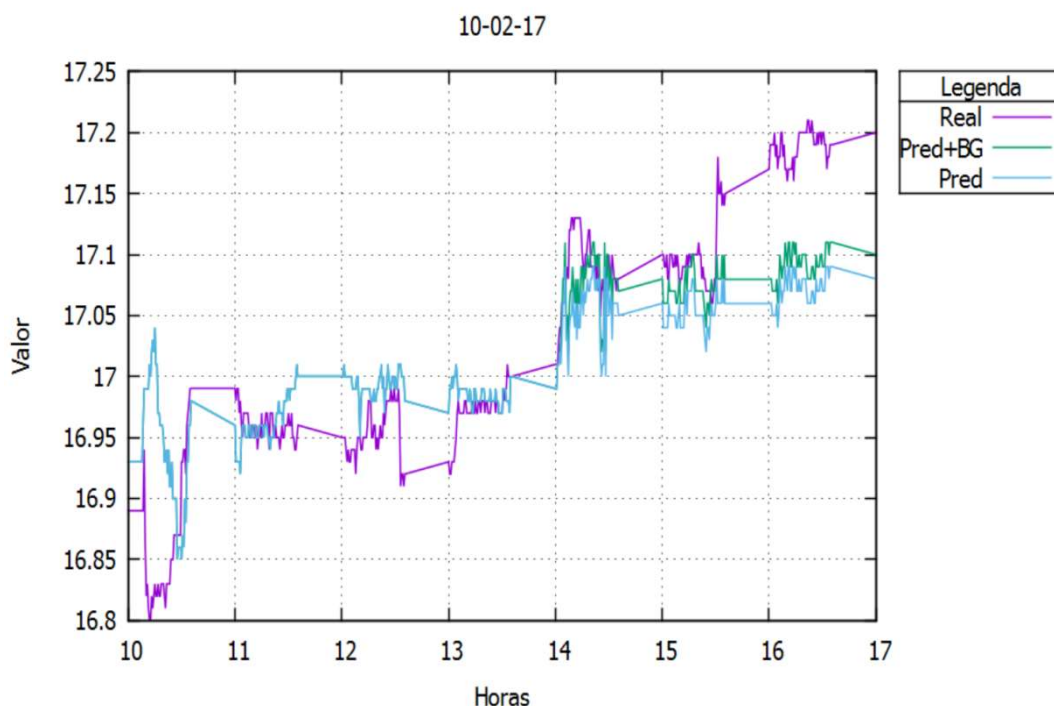
ação começou a subir e a predição apresentou um valor mais baixo que o preço da ação havendo um aumento do volume de negociações.

Nesse momento, o volume das ações foi maior que a média do volume, fazendo com que, nesse período, fossem acrescentados R\$0,03 representado pela linha verde.

Até às 14h, como o volume de ações negociadas não ultrapassou a média de volume do período, pela condição apresentada na Figura 28, fez sentido o *Big Data* não acrescentar os valores da aferição antes do volume atual ser mais alto que a média histórica.

A performance do *Big Data Hadoop* com um computador mestre e um escravo foi suficiente para gerar as estatísticas do período analisado e dentro do intervalo proposto, a cada 15 minutos.

Fig. 24 - Gráfico do dia 10/02/2017 apresentando o ajuste de R\$ 0,03 à predição depois de verificado a média do volume de ações.



Fonte: Criado pelo autor.

- Total de textos analisados: 21. 778.
- Total de textos classificados corretamente: 10.527.

- Porcentual de textos classificados corretamente: 48,34%.
- Total de textos classificados incorretamente: 11.251.
- Porcentual de textos classificados incorretamente: 51,66%.
- Média de acréscimo ao valor da predição: R\$ 0,02.

4.3 Resultado

A predição com *Naive Bayes* com 224 dias de coletas comparado ao histórico dos valores reais do preço da ação Abev3 conseguiu ficar próxima ao valor real. Em média, permaneceu no período analisado em R\$ 0,07 abaixo do valor real.

Pelo gráfico da Figura 26, a predição acompanhou, de forma macro, os picos positivos e negativos das oscilações do período.

O *Big Data* e a classificação dos textos financeiros com o processamento de linguagem natural obteve uma média de acertos de 48,34 %.

No período analisado, o resultado da classificação de textos foi de sentimentos positivos e, como média geral, foram adicionados R\$ 0,02 ao valor da predição, fazendo com que o resultado final da predição se aproximasse mais do valor real da ação.

5 CONCLUSÃO

O *Big Data* é ineficiente para processar grandes quantidades de pequenos arquivos de até 15 *kbytes*. Vários arquivos podem ser concatenados em um único arquivo com a função *merge* do *Hadoop*. Processando um único arquivo, haverá um ganho de velocidade.

Alguns sites financeiros em que as notícias foram extraídas estavam demorando mais de quinze minutos para atualização, o que prejudicou o dinamismo do sistema e o ajuste do preço real das ações pela classificação de textos.

Existe uma relação entre volume das ações negociadas e oscilação de preços. Normalmente, há uma tendência de subida do preço da ação quando existe um aumento repentino no volume negociado.

Apesar das diferentes formas de escritas dos autores que atualizavam os sites financeiros, a extração de textos com o processamento de linguagem natural identificou corretamente 48,34% dos textos analisados.

A média de ajuste de preços da predição com a classificação de textos foi de R\$ 0,02 positivos, aproximando-se do valor real.

A predição com *Naive Bayes* no período analisado, de maneira geral, conseguiu acompanhar, de forma macro, os picos positivos e negativos das oscilações dos valores e manteve-se em média R\$ 0,07 abaixo do valor real.

Os resultados das melhores predições, aquelas que se aproximaram mais do valor real, aconteceram no segundo bloco de medição, onde foram medidos 2/3 dos dados históricos dos preços das ações e não no terceiro bloco, onde foi analisado todo o conjunto de dados. Dessa forma, neste experimento, uma maior quantidade de dados históricos não necessariamente contribuiu para um melhor resultado.

O sistema cumpriu o objetivo fornecendo uma predição próxima do valor real. O *Big Data* e processamento de linguagem natural ajudou a melhorar a predição, aproximando-se do valor real. Não existem certezas de que um histórico sempre se repita e cada dia é um gráfico diferente.

Apesar desse experimento apresentar a pesquisa utilizando apenas uma empresa, esse estudo pode ser aplicado a outras empresas integrantes da bolsa de valores ou, até mesmo, de outros segmentos.

A decisão de investimento é sempre do investidor.

5.1 Trabalhos Futuros

Como sugestão para trabalhos futuros, que podem ser subsidiados por este estudo:

- a) **Validador de notícias falsas.** Este trabalho limitou-se a extrair as informações financeiras de sites de renomes. Nesse processo de extração, não foi verificada a veracidade da informação. É possível, portanto, aplicar um validador de notícias falsas, aumentando a credibilidade nas notícias que são extraídas.
- b) **Análise das bolsas e notícias mundiais.** Alguns incidentes que ocorrem em um ou mais países podem afetar as bolsas mundiais como tragédias, atentados, guerras, etc. Nesse contexto, a análise de outras bolsas, juntamente com o Ibovespa, pode ajudar a prever ou mesmo antecipar ações defensivas para o investidor.
- c) **Big Data na nuvem.** Hoje existe uma grande oferta de serviços de *Big Data* na Nuvem. Utilizando-se de uma infraestrutura de um *Data Center*, os benefícios das virtualizações, recursos computacionais sob demanda e disponibilidade 24 horas e com custos baixos, podem ser mais interessantes que manter uma infraestrutura interna de *Big Data*.
- d) **Machine Learning.** Ferramentas de aprendizado de máquina como o *framework Apache Mahout* e outros podem ser aplicados para estudo de comportamento do mercado e previsões auxiliando um investidor na tomada de decisões.

6 REFERÊNCIAS

APACHE HADOOP. **Consulta geral a homepage oficial.** Disponível em: <<http://hadoop.apache.org/>>. Acesso em 07/03/2016.

BULKOWSKY THOMAS N. **Encyclopedia of Chart Patterns**, 2nd ed. US. 2005. p.30.

CHAN KEITH C.C.**Big Data Analytics for Drug Discovery.**IEEE International Conference on Bioinformatics and Biomedicine. Hong Kong.2013.

CHOPRA D., MATHUR I., NISHEETH J., **Mastering Natural Language Processing with Python.** 1. ed. Packt Publishing Ltd. Birmingham B3 2PB, UK - Mumbai. June 2016.

CORREA, S. M. B. B., **Probabilidade e Estatística.** 2. ed. Pontifícia Universidade Católica de Minas Gerais. PUC Minas, 2003. Pag.8,9.

ELDER A., **Como se transformar em um operador e investidor de sucesso.** 15. ed. Elsevier. São Paulo, 2004. p. 47-52

GHOFRANI F., Haddad A. K., Jamshidi A. **Internet Traffic Classification Using Multiple Classifiers.** IEEE 7th International Conference on Information and Knowledge Technology. 2015. p. 1-5

GRAHAM BENJAMIM. O investidor inteligente. Rio de Janeiro, Nova Fronteira, 2015. p.32

Kim., Y, S. Jeong R, I. Ghani, **Text Opinion Mining to Analyze News for Stock Market Prediction.** Int. J. Advance. Soft Comput. Appl., Vol. 6, No. 1, March 2014

LEUNG C. K., JIANG F., **A Data Science Solution for Mining Interesting Patterns from Uncertain Big Data.** IEEE Fourth International Conference on Big Data and Cloud Computing, Sydney, NSW 2014. p.3-5

LEMOES FLÁVIO, **Análise Técnica dos Mercados Financeiros.** São Paulo, Saraiva, 2015.

PERKINS J., **Python Text Processing with NLTK 2.0 Cookbook.** 1. ed. Packt Publishing Ltd. Birmingham B3 2PB, UK - Mumbai. November 2010. p. 18

MANIKANDAN, S.G. AND RAVI, SIDDART H. **Big Data Analysis using Apache Hadoop**. IT Convergence and Security (ICITCS), International Conference. San Francisco, USA, 2015. p.700-703.

MATTAR, F. N. **Pesquisa de Marketing**. São Paulo, Atlas, 1994.

MOORE, MARCOS. **Ações: Quais comprar e onde comprar**. Rio de Janeiro, Campus, 2012. p. 54-60.

NUNZIO G.M., SORDONI A. **A Visual Tool for Bayesian Data Analysis: The Impact of Smoothing on Naïve Bayes Text Classifiers**. 35th Annual SIGIR Conference. Portland, Oregon, USA. 2012

OLIVEIRA EULINA., **Mercados reagem a BC dos EUA e Coreia do Norte**. São Paulo 09/09/2016 - 11:52.

<http://www1.folha.uol.com.br/mercado/2016/09/1811681-mercados-reagem-a-bc-dos-eua-e-coreia-do-norte-dolar-sobe-2-e-bolsa-cai-2.shtml>

PENA, S. D. **Thomas Bayes o cara**. Minas Gerais: Universidade Federal de Minas Gerais, 2006. Disponível em:

<http://www.icmc.usp.br/pessoas/francisco/sme0120/material/Thomas_Bayes_CH.pdf>. Acesso em 09/03/2016.

PRAJAPATI V., **Big Data Analytics With R and Hadoop**. 1. ed. Packt Publishing Ltd. Birmingham - Mumbai June. 2013

RAO D, GUDIVADA VENKAT, RAGHAVAN VIJAY, **Data Quality Issues in Big Data**. Big Data Conference. IEEE International Conference on Big Data. North Carolina USA. 2015. p.2654-2660.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>

SANTIDHANYAROJ P., KHAN T. A., GELOWITZ C. M., BENEDICENTI L., **A Sentiment Analysis Prototype System for Social Network Data**. Canadian Conference on Electrical and Computer Engineering. CCECE 2014

SHIHAVUDDIN., A. S. M., **Prediction of Stock Price analyzing the online financial news using Naive Bayes classifier and local economic trends**.

Advanced Computer Theory and Engineering, Volume:4. 3rd International Conference, Chengdu. 2010. p.V4-22-V4-26

SHYNKEVICH Y., **Stock Price Prediction based on Stock-Specific and Sub-Industry-Specific News Articles**. IEEE Neural Networks (IJCNN), 2015 International Joint Conference, Killarney. 2015.p 1-8.

YEJAS., O. D., ZHUANG W., PANNU A., **Big R: Large-Scale Analytics on Hadoop Using R**. Big Data (BigData Congress), IEEE International Conference. San Jose, CA, USA. 2014. p. 570-577

7 ANEXOS

Abaixo é apresentado a codificação do Map.py e Reduce.py.

```
map.py
#!/usr/bin/env python
#coding: utf-8

import sys
for texto in sys.stdin:
    minusculo = texto.lower() ##transforma em minúsculos
    print(minusculo);
```

Codificação do *Map.py* desenvolvido em *Python* e executado no *Hadoop* para classificação das palavras.

Fonte: Criado pelo autor.

```
#!/usr/bin/python
# -*- coding: utf-8 -*-

from operator import itemgetter
import sys
import string
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from unicodedata import normalize
from nltk.stem import WordNetLemmatizer
from nltk.stem import WordNetLemmatizer
from collections import Counter
from collections import Counter
import re
import nltk
import sys
import os
import datetime

contAlta = 0
contBaixa = 0

for tex3 in sys.stdin:
    # print(tex3)
    removePontuacao = re.sub('[%s]' % re.escape(string.punctuation), "", tex3) # remove
    pontuações
    removeNumeros = ".join([i for i in removePontuacao if not i.isdigit()]) # remove
    números
    removeNumerosU = unicode(removeNumeros, "utf-8")
    retiraAcentos = normalize('NFKD', removeNumerosU).encode('ASCII',
    'ignore').decode('ASCII')
    filtroStopWords = [w for w in word_tokenize(retiraAcentos) if not w in
    stopwords.words('portuguese')] # StopWords
    filtroStopWordsT = str(filtroStopWords)
    lemmatizer = WordNetLemmatizer()
    filtroLemma = lemmatizer.lemmatize(filtroStopWordsT) # returns u'dog'
    adjetivo = nltk.pos_tag(filtroLemma.split())
    tokens = nltk.word_tokenize(filtroLemma)
    pos_tagged_tokens = nltk.pos_tag(tokens)
    removePalavrasDuplicadasSTR = str(filtroLemma)
    palavrasChave = removePalavrasDuplicadasSTR.find('industrial') # bovespa
    counts = Counter(tag for word, tag in pos_tagged_tokens)
```

Codificação do *Reduce.py* desenvolvido em *Python* e executado no *Hadoop* para classificação das palavras.

Fonte: Criado pelo autor

```

dicionarioPositivo = [{"u'alta", 'JJ'}, {"u'aumenta", 'JJ'}, {"u'maximo", 'JJ'}, {"u'muito",
'JJ'}, {"u'positivo", 'JJ'}, {"u'valorizacao", 'JJ'}, {"u'abri", 'JJ'}, {"u'ascensao", 'JJ'}, {"u'ganha",
'JJ'}, {"u'dispara", 'JJ'}, {"u'forte", 'JJ'}, {"u'progreda", 'JJ'}, {"u'acelera", 'JJ'}, {"u'entra",
'JJ'}, {"u'mais", 'JJ'}, {"u'maior", 'JJ'}, {"u'melhor", 'JJ'}, {"u'sobe", 'JJ'}, {"u'estavel",
'JJ'}, {"u'soma", 'JJ'}, {"u'acredita", 'JJ'}, {"u'alegre", 'JJ'}, {"u'certo", 'JJ'}, {"u'abertura",
'JJ'}, {"u'progressivo", 'JJ'}, {"u'comemora", 'JJ'}, {"u'acerto", 'JJ'}, {"u'inclui", 'JJ'}, {"u'celebra",
'JJ'}, {"u'acumula", 'JJ'}, {"u'lidera", 'JJ'}, {"u'comemora", 'JJ'}, {"u'desconfia",
'JJ'}, {"u'beneficia", 'JJ'}, {"u'bem", 'JJ'}, {"u'alonga", 'JJ'}, {"u'entrada", 'JJ'}, {"u'acima",
'JJ'}, {"u'levanta", 'JJ'}, {"u'salta", 'JJ'}, {"u'pula", 'JJ'}, {"u'avanca", 'JJ'}, {"u'aumento",
'JJ'}, {"u'bom", 'JJ'}, {"u'alta", 'NN'}, {"u'aumenta", 'NN'}, {"u'maximo", 'NN'}, {"u'muito",
'NN'}, {"u'positivo", 'NN'}, {"u'valorizacao", 'NN'}, {"u'abri", 'NN'}, {"u'ascensao",
'NN'}, {"u'ganha", 'NN'}, {"u'dispara", 'NN'}, {"u'forte", 'NN'}, {"u'progreda", 'NN'}, {"u'acelera",
'NN'}, {"u'entra", 'NN'}, {"u'mais", 'NN'}, {"u'maior", 'NN'}, {"u'melhor", 'NN'}, {"u'sobe",
'NN'}, {"u'estavel", 'NN'}, {"u'soma", 'NN'}, {"u'acredita", 'NN'}, {"u'alegre", 'NN'}, {"u'certo",
'NN'}, {"u'abertura", 'NN'}, {"u'progressivo", 'NN'}, {"u'comemora", 'NN'}, {"u'acerto",
'NN'}, {"u'inclui", 'NN'}, {"u'celebra", 'NN'}, {"u'acumula", 'NN'}, {"u'lidera",
'NN'}, {"u'comemora", 'NN'}, {"u'desconfia", 'NN'}, {"u'beneficia", 'NN'}, {"u'bem",
'NN'}, {"u'alonga", 'NN'}, {"u'entrada", 'NN'}, {"u'acima", 'NN'}, {"u'levanta", 'NN'}, {"u'salta",
'NN'}, {"u'pula", 'NN'}, {"u'avanca", 'NN'}, {"u'aumento", 'NN'}, {"u'bom", 'NN'}]

dicionarioNegativo = [{"u'baixa", 'JJ'}, {"u'diminui", 'JJ'}, {"u'minimo", 'JJ'}, {"u'pouco",
'JJ'}, {"u'negativo", 'JJ'}, {"u'desvalorizacao", 'JJ'}, {"u'fecha", 'JJ'}, {"u'queda", 'JJ'}, {"u'perde",
'JJ'}, {"u'despenca", 'JJ'}, {"u'fraco", 'JJ'}, {"u'recua", 'JJ'}, {"u'frea", 'JJ'}, {"u'sai", 'JJ'}, {"u'menos",
'JJ'}, {"u'menor", 'JJ'}, {"u'pior", 'JJ'}, {"u'desce", 'JJ'}, {"u'instavel", 'JJ'}, {"u'subtrai",
'JJ'}, {"u'desconfia", 'JJ'}, {"u'triste", 'JJ'}, {"u'incerto", 'JJ'}, {"u'falencia", 'JJ'}, {"u'regressivo",
'JJ'}, {"u'chora", 'JJ'}, {"u'erro", 'JJ'}, {"u'retira", 'JJ'}, {"u'entristece", 'JJ'}, {"u'escasso",
'JJ'}, {"u'atras", 'JJ'}, {"u'lamenta", 'JJ'}, {"u'acredita", 'JJ'}, {"u'prejudica", 'JJ'}, {"u'mal",
'JJ'}, {"u'encurta", 'JJ'}, {"u'saida", 'JJ'}, {"u'abaixo", 'JJ'}, {"u'cai", 'JJ'}, {"u'afunda",
'JJ'}, {"u'rebaixa", 'JJ'}, {"u'retrocede", 'JJ'}, {"u'desconto", 'JJ'}, {"u'ruim", 'JJ'}, {"u'baixa",
'NN'}, {"u'diminui", 'NN'}, {"u'minimo", 'NN'}, {"u'pouco", 'NN'}, {"u'negativo",
'NN'}, {"u'desvalorizacao", 'NN'}, {"u'fecha", 'NN'}, {"u'queda", 'NN'}, {"u'perde",
'NN'}, {"u'despenca", 'NN'}, {"u'fraco", 'NN'}, {"u'recua", 'NN'}, {"u'frea", 'NN'}, {"u'sai",
'NN'}, {"u'menos", 'NN'}, {"u'menor", 'NN'}, {"u'pior", 'NN'}, {"u'desce", 'NN'}, {"u'instavel",
'NN'}, {"u'subtrai", 'NN'}, {"u'desconfia", 'NN'}, {"u'triste", 'NN'}, {"u'incerto",
'NN'}, {"u'falencia", 'NN'}, {"u'regressivo", 'NN'}, {"u'chora", 'NN'}, {"u'erro", 'NN'}, {"u'retira",
'NN'}, {"u'entristece", 'NN'}, {"u'escasso", 'NN'}, {"u'atras", 'NN'}, {"u'lamenta",
'NN'}, {"u'acredita", 'NN'}, {"u'prejudica", 'NN'}, {"u'mal", 'NN'}, {"u'encurta", 'NN'}, {"u'saida",
'NN'}, {"u'abaixo", 'NN'}, {"u'cai", 'NN'}, {"u'afunda", 'NN'}, {"u'rebaixa", 'NN'}, {"u'retrocede",
'NN'}, {"u'desconto", 'NN'}, {"u'ruim", 'NN'}]

for w in pos_tagged_tokens:
    contAlta += dicionarioPositivo.count(w)
    contBaixa += dicionarioNegativo.count(w)

```

Codificação do *Reduce.py* desenvolvido em *Python* e executado no *Hadoop* para classificação das palavras (continuação...)

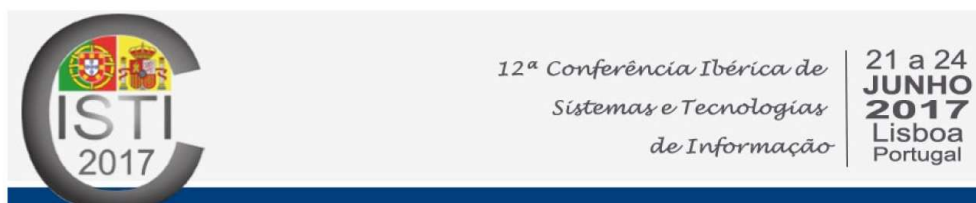
Fonte: Criado pelo autor

```
print(contAlta),  
print(contBaixa),  
print(len(pos_tagged_tokens))  
n1 = str(contAlta)  
n2 = str(contBaixa)  
n3 = str(len(pos_tagged_tokens))
```

Codificação do *Reduce.py* desenvolvido em *Python* e executado no *Hadoop* para
classificação das palavras (continuação...)

Fonte: Criado pelo autor

Abaixo os certificados e publicações em congressos em que o artigo deste trabalho foi submetido e aprovado.



Certificate

André Hideo Hayashi

presented the article

Big Data e computação preditiva para suporte a decisões no mercado financeiro

in the 12th Iberian Conference on Information Systems and Technologies (CISTI'2017), promoted
by Iberian Association for Information Systems and Technologies (AISTI), ISCTE-IUL and IEEE SMC
between the 21th and 24th of June 2017, at Lisbon, Portugal.



General Chair

**Certificado de apresentação no Congresso Internacional CISTI 2017 – Conferência
Ibérica de Sistemas e Tecnologias de Informação**



Certificate

André Hideo Hayashi

participated in the 12th Iberian Conference on Information Systems and Technologies (CISTI'2017), promoted by Iberian Association for Information Systems and Technologies (AISTI), ISCTE-IUL and IEEE SMC between the 21th and 24th of June 2017, at Lisbon, Portugal.



General Chair

Certificado de participação no Congresso Internacional CISTI 2017 – Conferência Ibérica de Sistemas e Tecnologias de Informação

ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=André%20Hideo

IEEE.org | IEEE Xplore Digital Library | IEEE-SA | IEEE Spectrum | More Sites

Cart (0) | Create Account | Personal Sign In

IEEE Xplore® Digital Library

Institutional Sign In

IEEE

BROWSE ▼ MY SETTINGS ▼ GET HELP ▼ WHAT CAN I ACCESS? SUBSCRIBE

André Hideo

Search

Basic Search Author Search Publication Search Advanced Search Other Search Options ▼

Displaying 1 of 1 result for **André Hideo** ✕

☐ Select All on Page Export ▼ Set Search Alerts ▼ Search History

☐ **Big data and predictive computing to support financial market decisions** André Hideo Hayashi

2017 12th Iberian Conference on Information Systems and Technologies (CISTI)

Year: 2017

Pages: 1 - 6, DOI: 10.23919/CISTI.2017.7975837

IEEE Conference Publications

Abstract (399 Kb) ©

Need Full-Text access to IEEE Xplore for your organization? REQUEST A FREE TRIAL >

IEEE Xplore®

Publicação do artigo no site IEEE Xplore



Certificado de apresentação na ERIPI 2017 – Escola Regional da Informática do Piauí



Certificado de participação na ERIPI 2017 – Escola Regional da Informática do Piauí