UNIVERSIDADE FUMEC
FACULDADE DE CIÊNCIAS EMPRESARIAS - FACE

Freise Wanderson Gonçalves de Andrade

MINERAÇÃO DE TEXTOS: IDENTIFICANDO TENDÊNCIAS NO MERCADO DE AÇÕES POR MEIO DOS SENTIMENTOS EXTRAÍDOS DE NOTÍCIAS PUBLICADAS NA INTERNET

Belo Horizonte
2018

Freise Wanderson Gonçalves de Andrade

MINERAÇÃO DE TEXTOS: IDENTIFICANDO TENDÊNCIAS NO MERCADO DE AÇÕES POR MEIO DOS SENTIMENTOS EXTRAÍDOS DE NOTÍCIAS PUBLICADAS NA INTERNET

Dissertação apresentada ao Programa de Pós-Graduação *Stricto Sensu* em Sistemas de Informação e Gestão de Conhecimento da Universidade FUMEC como requisito parcial para a obtenção do título de Mestre em Sistemas de Informação e Gestão de Conhecimento.

Área de concentração: Gestão de Sistemas de Informação e do Conhecimento.

Linha de Pesquisa: Sistemas e Tecnologia de Informação.

Orientador: Prof. Dr. Luiz Cláudio Gomes Maia.

Belo Horizonte-MG

2018

Dados Internacionais de Catalogação na Publicação (CIP)

A553m Andrade, Freise Wanderson Gonçalves de, 1981 -

Mineração de textos: identificando tendências no mercado de ações por meio dos sentimentos extraídos de notícias publicadas na internet / Freise Wanderson Gonçalves de Andrade. – Belo Horizonte, 2018.

104 f : il. ; 29,7 cm

Orientador: Luiz Cláudio Gomes Maia

Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento), Universidade FUMEC, Faculdade de Ciências Empresariais, Belo Horizonte, 2018.

1. Aprendizado do computador - Brasil. 2. Emoções - Análise. 3. Mercado de ações - Previsão - Brasil. I. Título. II. Maia, Luiz Cláudio Gomes. III. Universidade FUMEC, Faculdade de Ciências Empresariais.

CDU: 336.76

Ficha catalográfica elaborada pela Biblioteca Universitária-FUMEC

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Luiz Cláudio Gomes Maia, que me aconselhou e ajudou no desenvolvimento deste trabalho.

Aos professores e amigos que fiz no Mestrado em Sistemas de Informação e Gestão do Conhecimento que me possibilitaram chegar até aqui.

A minha família, que esteve ao meu lado em todos os momentos.

A todos, muito obrigado!

RESUMO

No mercado contemporâneo de ações, notícias publicadas sobre determinada empresa podem influenciar o preço de suas ações na bolsa de valores, ainda que de forma rápida e meramente especulativa. Dessa forma, um investidor tentando minimizar perdas e maximizar ganhos procura informações referentes a empresas das quais possui ações em diversas fontes como *sites* de notícias, balanços, gráficos, relatórios, entre outras, porém, essa tarefa pode tomar muito tempo ao ser executada por uma pessoa. Com isso a utilização de sistemas de informação torna-se essencial para automatizar essas atividades. Esta pesquisa teve como objetivo avaliar a precisão de um sistema de informação para identificar a tendência de baixa ou alta do papel PETR4 da empresa Petróleos do Brasil S/A. (PETROBRAS) na Bolsa de Mercadorias e Futuros da Bolsa de Valores de São Paulo (BMF&BOVESPA) por meio da extração dos sentimentos dos *feeds* de notícias publicados na internet envolvendo a empresa. Esses sentimentos foram utilizados como características para os classificadores de aprendizado de máquina. O método de pesquisa seguiu o *Design Science Research* como forma de atenuar o distanciamento entre teoria e prática. No desenvolvimento da pesquisa foi possível medir o sentimento expressado por meio das notícias publicadas na internet e verificou-se que, apesar de as notícias terem a capacidade de influenciar as ações dos investidores e com isso exercerem efeitos importantes sobre o mercado de ações, avaliar o movimento do mercado somente por meio dos sentimentos carregados nas notícias mostrou-se ser uma tarefa complexa. Utilizando-se do *corpus* de notícias construído no desenvolvimento deste trabalho foi obtida precisão de 72% na identificação de tendências do papel PETR4 utilizando o classificador *Naive Bayes*. Apesar, porém, dos 72% de precisão na identificação de tendências alcançados no experimento, percebe-se que a abordagem para utilizar os sentimentos carregados pelas notícias enfrenta uma série de desafios, como compreender os fundamentos e a variação do impacto que essas notícias exercem sobre os investidores ao longo do tempo.

Palavras-chave: Mineração de Texto. Aprendizado de Máquina. Análise de Sentimentos. Mercado de ações.

ABSTRACT

In the contemporary stock market, news about a particular company can influence the price of their shares on the stock market, albeit quickly and merely speculatively. In this way an investor trying to minimize losses and maximize gains looks for information about companies that have actions in various sources such as news sites, balance sheets, charts, reports among others, but this task can take a lot of time to be executed by a person, with this the use of information systems becomes essential to automate these activities. This research aimed to evaluate the accuracy of an information system to identify the low or high tendency of Petróleos do Brasil S/A. (PETROBRAS) in the Bolsa de Mercadorias e Futuros da Bolsa de Valores de São Paulo (BMF&BOVESPA) through the extraction of the feelings of news feeds published on the Internet involving the company, these feelings were used as characteristics for machine learning classifiers. The research method followed the Design Science Research method as a way to mitigate the gap between theory and practice. In the development of the research it was possible to measure the sentiment expressed through the news published on the internet and it was verified that although the news had the capacity to influence the actions of the investors and with that they had important effects on the stock market, to evaluate the movement of the only through the feelings loaded in the news proved to be a complex task. Using the corpus of news built in the development of this work, a precision of 72% was obtained in the identification of PETR4 paper trends using the Naive Bayes classifier. However, despite the 72% accuracy of trend identification achieved in the experiment, it is perceived that the approach to use the feelings charged by the news faces a series of challenges such as understanding the fundamentals and the variation of the impact that this news has on investors over of time.

Keywords: Text Mining, Machine Learning, Sentiment Analysis, Stock Market.

LISTA DE SIGLAS

ANN	Redes Neurais Artificiais
API	*Application Programming Interface*
ARFF	*Attribute Relation File Format*
BMF&BOVESPA	Bolsa de Mercadorias e Futuros da Bolsa de Valores de São Paulo
CEP	Código de endereçamento postal
CRM	Relações com o cliente
DSR	*Design Science Research*
DSRM	*Design Science Research Methodology*
FUMEC	Fundação Mineira de Educação e Cultura
GNU	*General Public License*
HTML	*HyperText Markup Language*
IA	Inteligência artificial
IaaS	Infraestrutura como serviço
IBM	*International Business Machines*
JSON	*JavaScript Object Notation*
k-NN	*k-Nearest Neighbors*
PaaS	Plataforma como serviço
PETROBRAS	Petróleos do Brasil S/A.
PLN	Processamento de Linguagem Natural
REST	*Representational State Transfer*
RI	Recuperação da informação
RSS	*Really Simple Syndication*
SaaS	*Software* como serviço
SFI-ASM	*Santa Fe Institute Artificial Stock Market*
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency–Inverse Document Frequency*
URL	*Uniform Resource Locator*
www	*World wide web*
XML	*Extensible Markup Language*

LISTA DE FIGURAS

Figura 1- Arquitetura de um sistema genérico de mineração de texto 23
Figura 2 - Arquitetura em alto nível de um sistema de recuperação da informação . 27
Figura 3 - Metodologia para realizar a aprendizagem de máquina 31
Figura 4 - Exemplo de retorno de uma REST API no formato JSON 37
Figura 5 - Exemplo de retorno de uma REST API no formato XML 38
Figura 6- Processamento do texto por meio da API do IBM *Watson Tone Analyzer* 42
Figura 7 - Processo de seleção dos artigos na revisão sistemática da literatura 48
Figura 8 - Artigos por ano 52
Figura 9 - *Design Science Research Methodology* 61
Figura 10 - Modelo de entidades 64
Figura 11 - Perfil da empresa Petrobras 66
Figura 12 - Interface de cadastro e edição das empresas 67
Figura 13 - Exemplo de um documento RSS simples 68
Figura 14 - Interface de cadastro e edição das URLs dos endereços RSS dos *feeds* de notícias. 70
Figura 15 - Componente de extração e armazenamento dos *feeds* de notícias 71
Figura 16 - Extração dos sentimentos dos *feeds* de notícias por meio da API do *IBM Watson Tone Analyzer* 74
Figura 17 - Exemplo de um arquivo no formato ARFF 75
Figura 18 - *View* resultado da união dos dados da tabela *Quotation* e IdxIBM*WatsonToneAnalyze*r 76
Figura 19 - Faixa de probabilidade de um índice estar correto definido pela API do IBM *Watson Tone Analyzer* 77
Figura 20 - Precisão 79
Figura 21 - Fórmula para calcular a precisão 79
Figura 22 - Revocação 80
Figura 23 - Fórmula para calcular a revocação 80
Figura 24 - Matriz confusão 81
Figura 25 - Modelo arquitetural em alto nível da aplicação do experimento 82
Figura 26 - Resultado da matriz confusão utilizando o classificador KNN 83
Figura 27 - Resultado da matriz confusão utilizando o classificador *Naive Bayes* 84

Figura 28 - Resultado da matriz confusão utilizando o classificador SVM 84
Figura 29 - Tela inicial para acesso aos artefatos desenvolvidos 92

LISTA DE GRÁFICOS

Gráfico 1 - Número de negócios realizados na BM&FBovespa por meio de robôs investidores em %.....34
Gráfico 2 - Distribuição dos artigos encontrados na pesquisa automática por base de busca.....51
Gráfico 3 - Técnicas de pré-processamento mais utilizadas.....54
Gráfico 4 - Algoritmos mais utilizados.....56
Gráfico 5 - As 10 maiores precisões por algoritmo.....57
Gráfico 6 - Precisão por tamanho do *corpus*.....58
Gráfico 7 - Quantidade de notícias extraídas por mês.....72
Gráfico 8 - Fechamento diário do papel PETR4 no período de 1/11/2016 a 31/10/2017.....73
Gráfico 9 - Comparação entre classificadores por precisão e revocação.....83
Gráfico 10 - Percentual de acertos entre a classe -1 e 1 do experimento 1.....85
Gráfico 11 - Fechamento do papel PETR4 x quantidade de notícias classificadas como negativas e positivas no mês.....86
Gráfico 12 - Soma mensal dos índices dos sentimentos e a relação com o fechamento mensal do papel PETR4.....87

LISTA DE TABELAS

Tabela 1 - Quantidade de artigos obtidos após aplicação dos critérios de inclusão .51
Tabela 2 - Artigos excluídos.....52
Tabela 3 - Notícias que foram classificadas com sentimento negativo, porém, o fechamento do papel PETR4 foi positivo no mercado de ações.....89
Tabela 4 - Notícias que foram classificadas com sentimento positivo, porém o fechamento do papel PETR4 foi negativo no mercado de ações 89

LISTA DE QUADROS

Quadro 1 - Questões de pesquisa da revisão sistemática da literatura 45
Quadro 2 - Atividades do processo proposto por Peffers 61
Quadro 3 - Descrição das tabelas do modelo de entidade e relacionamento 65
Quadro 4 - Descrição dos arquivos de treinamento gerados 75

SUMÁRIO[1]

1 INTRODUÇÃO 15
1.1 Contexto e relevância da temática 15
1.2 Lacuna a ser explorada 16
1.3 Problema de pesquisa 16
1.4 Objetivos 17
1.4.1 Geral *17*
1.4.2 Específicos *17*
1.4.3 Adequação à linha de pesquisa *17*

2 FUNDAMENTAÇÃO TEÓRICA 18
2.1 Análise de sentimentos 18
2.1.1 Análise de sentimentos e o mercado financeiro. 19
2.2 Mineração de textos 21
2.2.1 Coleção de documentos. 23
2.2.2 Etapas da mineração textos. 24
2.2.3 Recuperação de informação. 25
2.2.4 Processamento de linguagem natural. 28
2.3 Aprendizado de máquina 29
2.3.1 Aprendizado supervisionado. 30
2.3.2 Aprendizado não supervisionado. 31
2.4 Robôs operando no mercado financeiro 32
2.5 Computação em nuvem e APIs para identificação de sentimentos 35
2.5.1 Application Programming Interface 36
2.5.2 REST-API. 36
2.5.3 APIS para identificação de sentimentos. 39

3 TRABALHOS RELACIONADOS 43

[1] Este trabalho foi revisado de acordo com as novas regras ortográficas aprovadas pelo Acordo Ortográfico assinado entre os países que integram a Comunidade de Países de Língua Portuguesa (CPLP), em vigor no Brasil desde 2009. E foi formatado de acordo com a ABNT NBR 14724 de 17.04.2016.

4 REVISÃO SISTEMÁTICA DA LITERATURA 44
4.1 Introdução 44
4.2 Metodologia da revisão sistemática da literatura 44
4.2.1 Questões de pesquisa 45
4.2.2 Amplitude da pergunta 45
4.2.3 Planejamento da revisão sistemática 46
4.2.4 Critérios para seleção 47
4.2.4.1 Critério de inclusão. 47
4.2.4.2 Critério de exclusão. 47
4.2.4.3 Processo de seleção. 47
4.2.4.4 Realização das buscas. 49
4.2.5 Estratégias de extração de dados e sumarização dos resultados 50
4.3 Resultados da revisão sistemática da literatura 50
4.4 Conclusão da revisão sistemática da literatura 58

5 METODOLOGIA 60
5.1 Caracterização da pesquisa 60
5.2 A metodologia proposta por Peffers aplicada à pesquisa 60
5.3 Design e desenvolvimento. 62
5.3.1 Ferramentas utilizadas no desenvolvimento do trabalho *62*
5.3.2 Criação da base de dados *64*
5.3.3 Seleção da empresa para realização dos experimentos *65*
5.3.4 Seleção dos sites de notícias *67*
5.3.5 Extração e criação do corpus de notícias e cadastramento do fechamento diário do papel PETR4 ..... *71*
5.3.6 Pré-Processamento do corpus de notícias *73*
5.3.7 Geração do arquivo de treinamento *74*
5.3.7.1 Geração do arquivo de treinamento para o experimento. 76
5.4 Demonstração e avaliação do experimento 78
5.4.1 Técnicas de experimento e validação 81
5.5 Análise dos resultados 82
5.6 Comunicação e contribuições gerais desta pesquisa. 91

6 CONSIDERAÇÕES FINAIS 93
6.1 Limitação da pesquisa e trabalhos futuros 94

REFERÊNCIAS 96

1 INTRODUÇÃO

1.1 Contexto e relevância da temática

Antecipar o movimento do mercado de ações sempre foi uma tarefa desafiadora devido, à sua alta volatilidade e dinâmica. No mercado financeiro, as oportunidades precisam ser exploradas assim que surgem, sendo que o investidor que conseguir antecipar o movimento de um papel (alta ou baixa) terá vantagens em relação aos demais (DUONG; NGUYEN; DANG, 2016; GIDOFALVI, 2001). Dessa forma, um investidor tentando minimizar perdas e maximizar ganhos procura informações referentes às empresas das quais possui ações, em diversas fontes, como: notícias, balanços, gráficos, relatórios entre outras. Essa tarefa, porém, pode tomar muito tempo ao ser executada por uma pessoa, vista a quantidade de informações que devem ser analisadas. Torna-se mais crítica quando se leva em conta a informação que pode ser gerada e transmitida por meio da *Web*, em *blogs*, mídias das empresas e, principalmente, *sites* de notícias sobre o mercado financeiro (LOPES *et al.*, 2008).

Pesquisas na área de investimentos já averiguaram que no mercado moderno de ações as notícias publicadas podem influenciar o preço das ações na bolsa, ainda que de forma rápida e puramente especulativa. Quem determina esse valor é quem negocia no momento da compra e venda. Logo essa opinião global certamente está relacionada ao valor de um ativo no mercado. Então, teoricamente seria provável utilizando processos computacionais extrair as informações relevantes dessas fontes de notícias a fim de ajudar os investidores em sua tarefa de coletar informação que auxilie na tomada de decisão de um investimento (LOPES *et al.,* 2008; MAO; COUNTS; BOLLEN, 2011; MAO; ZENG, 2011).

As informações extraídas de notícias são portadoras de opiniões que poderão ser positivas, negativas ou neutras. Um conjunto de opiniões sobre determinada empresa de capital aberto com papéis na bolsa de valores pode alterar o estado emocional dos investidores naquele papel que, por questões emocionais, pode induzi-los a comprar, vender ou mantê-lo. Com isso, pode-se fazer uma correlação entre um conjunto de opiniões extraídas das notícias com o movimento de alta ou baixa de determinado papel pertencente a uma empresa (ACKERT; CHURCH;

DEAVES, 2003). Nesse cenário, sistemas de informações automatizados tornam-se importantes instrumentos para auxiliar investidores na identificação de tendências no mercado de ações, fornecendo subsídios a esses investidores para uma negociação mais assertiva.

1.2 Lacuna a ser explorada

O conteúdo dos meios de comunicação *online* tem se mostrado um importante fator que molda o sentimento dos investidores por meio do sentimento que o texto das notícias carrega (MAO; COUNTS; BOLLEN, 2011). A economia comportamental afirma que as emoções podem afetar profundamente o comportamento e a tomada de decisão de um indivíduo (BOLLEN; MAO; ZENG, 2011; NOFSINGER, 2005). Se a emoção do indivíduo investidor pode afetar a forma como ele reage às novas informações, é aceitável que o sentimento coletivo dos investidores possa influenciar a dinâmica do mercado de ações (OLIVEIRA; CORTEZ; AREAL, 2013). Dessa forma, realizando a mineração de texto em um conjunto de notícias extraídas de *sites* especializados no mercado financeiro, a extração dos sentimentos carregados por esse conjunto de notícias e a aplicação de técnicas de aprendizado de máquina, pode-se identificar a tendência de determinado papel no mercado de ações, o que pode justificar novos estudos nesse campo e o incentivo ao desenvolvimento de sistemas de informação que possam automatizar esse processo.

1.3 Problema de pesquisa

Qual a precisão de um sistema de informação para a identificação de tendências no mercado de ações por meio dos sentimentos extraídos de notícias publicadas na internet, utilizando técnicas de mineração de texto e aprendizado de máquina como ferramenta para auxiliar investidores na tomada de decisões?

1.4 Objetivos

1.4.1 Geral

O objetivo geral desta pesquisa foi avaliar a precisão de um sistema de informação para identificar a tendência do papel PETR4 da empresa Petróleos do Brasil S/A. (PETROBRAS) na Bolsa de Mercadorias e Futuros da Bolsa de Valores de São Paulo (BMF&BOVESPA) por meio da extração dos sentimentos dos *feeds* de notícias publicados na internet.

1.4.2 Específicos

Como objetivos específicos:

a) Descrever métodos e técnicas comumente empregados na identificação de tendências no mercado de ações por meio de notícias publicadas na Internet.
b) Construir um *corpus* de notícias financeiras sobre a Petrobras.
c) Elaborar algoritmos computacionais para extração das notícias e criação dos arquivos de treinamento que serão utilizados pelos algoritmos de aprendizado de máquina.

1.4.3 Adequação à linha de pesquisa

O Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC está estruturado em uma área de concentração: Gestão de Sistemas de Informação e do Conhecimento, abrangendo duas linhas de pesquisa, sendo: Gestão da Informação e do Conhecimento e Sistemas e Tecnologia de Informação.

Este trabalho posiciona-se na linha de pesquisa Tecnologia e Sistemas de Informação, que se encaixa no objetivo deste trabalho, que é avaliar a precisão de um sistema de informação para identificar a tendência do papel PETR4 da empresa Petrobras no mercado de ações, utilizando técnicas de mineração de texto, análise de sentimentos e classificadores de aprendizado de máquina.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda os principais conceitos que proveram subsídios para o desenvolvimento do trabalho, contemplando os seguintes tópicos: análise de sentimentos, análise de sentimentos e o mercado financeiro, mineração de textos, recuperação de informação, processamento de linguagem natural, aprendizado de máquina, robôs que operam no mercado financeiro e computação em nuvem e *Application Programming Interface* (APIs) para identificação de sentimentos.

2.1 Análise de sentimentos

A análise de sentimento, também chamada de mineração de opinião, é o campo de estudo que analisa as opiniões, sentimentos, avaliações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos e seus atributos. Possui também muitos outros nomes ligeiramente diferentes, como, por exemplo: extração de opinião, mineração de sentimento, análise de subjetividade, análise de afeto, análise de emoções, mineração de revisões, etc., no entanto, eles estão todos sob a análise de sentimento ou mineração de opinião. Enquanto nas organizações a análise de sentimento é o termo mais comumente utilizado, na academia tanto a análise de sentimentos e a mineração de opinião são frequentemente empregados, representando basicamente o mesmo campo de estudo (LIU, 2012).

A análise de sentimentos como uma subdisciplina da mineração de dados e linguística computacional é referida como as técnicas computacionais utilizadas para extrair, classificar, compreender e avaliar as opiniões expressas em várias fontes de notícias *online*, comentários de mídia social e outros conteúdos gerados (MARTINAZZO; PARAISO, 2010). A análise de sentimento é usada frequentemente para identificar sentimentos, afeto, subjetividade e outros estados emocionais em textos.

Pesquisas nessa área procuram demonstrar que a linguagem reflete mais do que se quer dizer. A frequência com que certos tipos de palavras são utilizados pode fornecer pistas sobre a personalidade, estilo de pensamento, conexões sociais e os estados emocionais. Como exemplo, tem-se o trabalho de Martinazzo e Paraiso

(2010), que apresentou um método de identificação de emoções em bases textuais em língua portuguesa. No caso dessa pesquisa, textos curtos foram manchetes de notícias diversas, extraídas de *sites* da internet, seguidas de uma breve descrição com o objetivo de identificar uma das seis emoções básicas descritas por Paul Ekman e Wallace Friesen (alegria, raiva, tristeza, desgosto, medo e surpresa) em notícias curtas.

2.1.1 Análise de sentimentos e o mercado financeiro

O mercado financeiro é responsável por fazer a ligação entre indivíduos e empresas com interesses em captar ou emprestar recursos financeiros para fins e motivos variados; é uma entidade que permeia a economia. De um lado, existe um grupo de poupadores, aqueles cuja renda lhes permite atender às suas necessidades de consumo imediato e, ainda, preservar parte para aplicar no mercado. De outro lado, estão aqueles que precisam de recursos suplementares para fazer frente às suas obrigações imediatas, sejam elas para consumo ou para investimentos produtivos (SANTOS; SANTOS, 2005). Existe um terceiro grupo participante que se pode chamar de intermediários, que são instituições que têm dinheiro para investir em pessoas ou empresas que demandam de fundos para financiar suas operações. Dessa forma, são os responsáveis por criar o mercado (HILLIER; GRINBLATT; TITMAN, 2011; SANTOS; SANTOS, 2005). O mercado financeiro se divide em dois grupos: o mercado monetário, composto pelos títulos de dívida a curto prazo, e o mercado de capitais, onde são negociados títulos de dívida de longo prazo e as ações e que compreende um grupo constituído de mediadores e instituições de apoio ao Sistema Financeiro Nacional que fazem convergir os interesses de tomadores e emprestadores, fazendo o capital movimentar pela economia.

O mercado de capitais no Brasil tem apresentado significativo crescimento nos volumes financeiros negociados desde a década de 2000. De acordo com especialistas de mercado, isso se deve principalmente à implementação de regras de governança claras e regras mais rígidas para a gestão de empresas e a entrada em vigor da Lei 10.303/01 no final de 2001. De fato, entre 1994 e 2001, o volume

médio diário negociado no mercado de ações aumentou 2,5 vezes. Entre 2001 e 2013 o volume foi multiplicado por 12 (SAITO; TULIO; PADILHA, 2013).

O mercado de ações é uma parte importante e ativa do mercado financeiro hoje em dia e tanto os investidores como os especuladores gostariam de obter melhores lucros analisando as informações sobre o mercado. Com isso, os artigos de notícias são reconhecidos como uma fonte importante de informações, amplamente utilizados e analisados por investidores. Na era do *Big Data* a quantidade de artigos de notícias tem aumentado tremendamente. Diante de um volume tão grande de notícias, mais e mais instituições dependem do alto poder de processamento dos computadores modernos para análise. As previsões dadas pelos sistemas de apoio poderiam ajudar os investidores a filtrar ruídos e tomar melhores decisões. Portanto, identificar novas formas de modelar e analisar artigos de notícias torna-se um problema interessante. Os artigos de notícia são primeiro interpretados por investidores e traduzidos em sentimento do mercado, em que os investidores, em seguida, tomam suas decisões com base no sentimento e suas interpretações. Com isso, os preços de mercado agregam as ações de cada investidor e refletem-nas no movimento do preço final (LI, X. *et al.*, 2014).

O valor de uma ação é constituído a partir das decisões dos investidores, não existe mercado sem investidores, são eles que durante suas práticas criam os preços das ações. Gunther (2002), em seu livro “Os axiomas de Zurique” destinado aos investidores, fala sobre esse assunto:

> A bolsa de valores, por exemplo, é um gigantesco mecanismo de emoções humanas. O que homens e mulheres estão fazendo, pensando, sentindo é que determina as altas e baixas das ações. O preço das ações de determinada empresa não sobe por causa dos dados abstratos num balancete, nem porque as perspectivas futuras da empresa são objetivamente boas. O mercado não desaba porque um computador, num canto qualquer, de algum modo, resolveu que está subindo a pressão vendedora, mas porque pessoas estão preocupadas, desanimadas ou temerosas. Ou simplesmente porque vem um fim de semana prolongado, quatro dias feriados e os compradores foram todos para a praia (GUNTHER, 2017, p. 65).

Os economistas cognitivos e comportamentais consideram o preço como um valor puramente percebido e não derivado do custo de produção. A mídia não relata apenas o *status* do mercado, mas elas criam ativamente um impacto na dinâmica do mercado com base nas notícias que lançam. Os participantes do mercado têm

vieses cognitivos, como excesso de confiança, reação exagerada, viés representativo, viés de informação e vários outros erros humanos previsíveis em raciocínio e processamento de informações.

A ciência comportamental e a teoria do sentimento dos investidores estabelecem de forma clara que o comportamento dos investidores pode ser moldado se eles se sentem otimistas ou pessimistas. A previsão do mercado de ações está intimamente relacionada ao clima dos participantes públicos ou do mercado, conforme estabelecido pela economia comportamental. No entanto, no caso da análise do sentimento em relação a um produto, a identificação do que um texto implica é muito mais direta do que no caso da previsão no mercado de ações. Em geral, a opinião de um produto implica emoções positivas ou negativas sobre ele, que para o mercado de ações pode não ser tão simples (KHADJEH NASSIRTOUSSI *et al.,* 2014).

2.2 Mineração de textos

A mineração de textos pode ser definida como um processo no qual um usuário interatua com uma coleção de documentos ao longo do tempo, usando um conjunto de ferramentas de análise. De forma equivalente à mineração de dados, a mineração de texto busca extrair informações úteis de fontes de dados por meio da identificação e exploração de padrões. No caso da mineração de texto, no entanto, as fontes de dados são coleções de documentos e os padrões são encontrados não entre registros de banco de dados formalizados, mas nos dados textuais não estruturados nos documentos dessas coleções (FELDMAN; SANGER, 2006).

No estudo realizado por Abdullah, Rahaman e Rahman (2013) com o objetivo de extrair informações fundamentais de fontes de notícias relevantes e usá-las para analisar o mercado de ações da Bolsa de Valores de Daca, Bangladesh, do ponto de vista do investidor comum, os autores propuseram uma nova estrutura de processamento de textos de diferentes fontes. Como entrada, as informações eram passadas por uma etapa de pré-processamento, utilizando-se uma ferramenta de processamento de linguagem natural para extrair a maior quantidade de informações possíveis e, portanto, definir uma classificação ou peso para essas informações, comparando com dados históricos. Segundo os autores, qualquer notícia pode levar

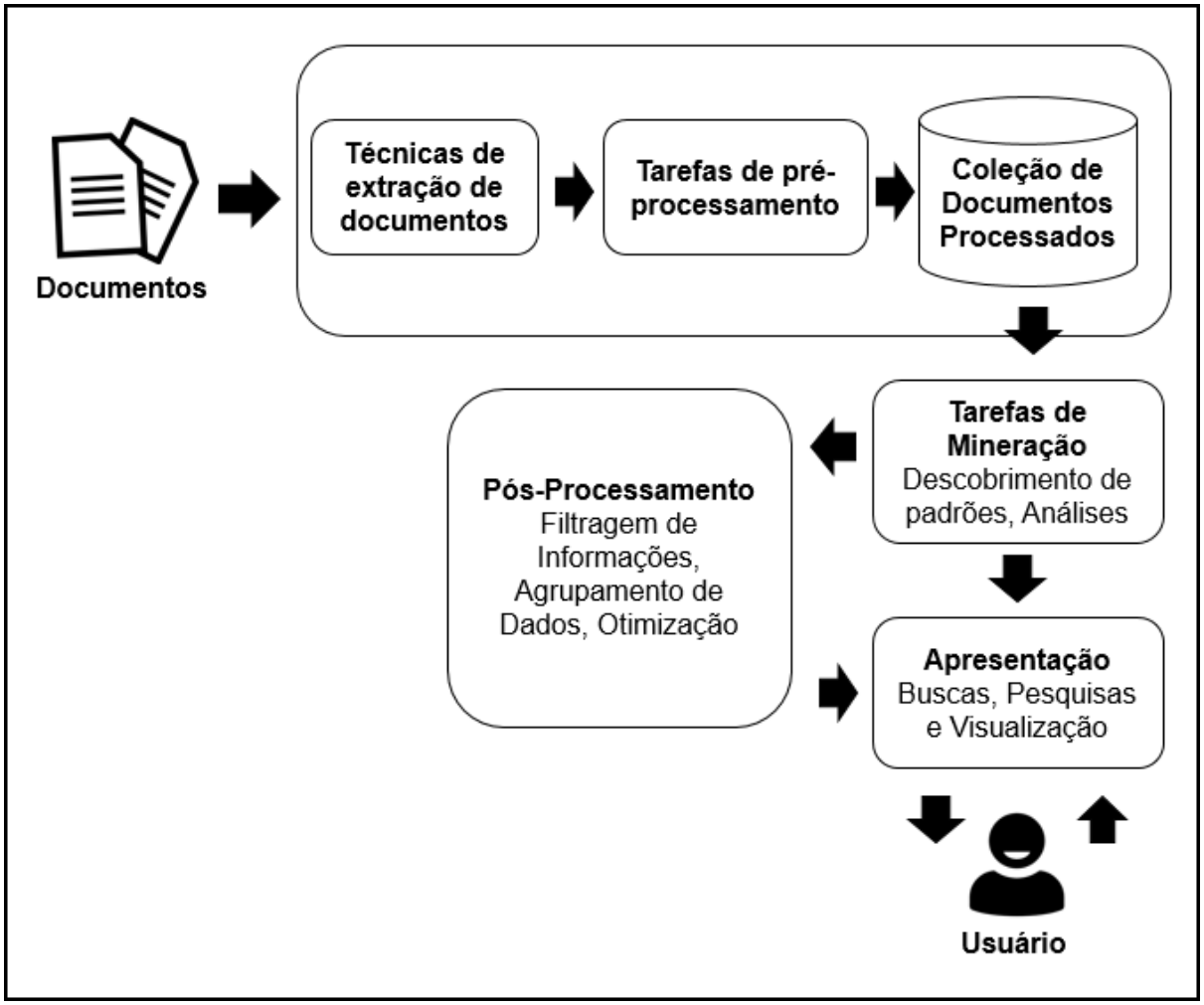
o investidor à tomada de decisão, comparando-a com dados cronológicos que podem ou não ser verdadeiros, dependendo de outros fatores, como tendências de preços.

A mineração de texto deriva grande parte da pesquisa seminal sobre mineração de dados. Para Feldman e Sanger (2006), portanto, não é surpreendente descobrir que os sistemas de mineração de texto e de mineração de dados confirmam muitas semelhanças arquiteturais de alto nível. Isso exemplifica o fato de que ambos os tipos de sistemas necessitam utilizar rotinas de pré-processamento, algoritmos de descoberta de padrões e elementos para a camada de apresentação como instrumentos de visualização para aprimorar a navegação do conjunto de respostas.

A mineração de texto adota tipos específicos de padrões em suas operações básicas de descoberta de informação que foram introduzidas e examinadas pela primeira vez na pesquisa de mineração de dados. Como a mineração de dados pressupõe que os dados já foram armazenados em um formato estruturado, seu esforço de pré-processamento concentra-se em duas tarefas críticas: normalizar dados e criar um grande número de associações entre tabelas, já para os sistemas de mineração de texto as operações de pré-processamento se preocupam na identificação e extração de características representativas para documentos em linguagem natural. Essas operações de pré-processamento são responsáveis pela transformação de dados não estruturados armazenados em coleções de documentos em um formato intermediário mais explicitamente estruturado, o que é uma preocupação que não é relevante para a maioria dos sistemas de mineração de dados (HU; LIU, 2004).

Segue na FIG. 1 o exemplo da arquitetura de um sistema genérico de mineração de texto.

Figura 1 - Arquitetura de um sistema genérico de mineração de texto



Fonte: adaptada de Feldman e Sanger (2006).

Devido à centralidade do texto de linguagem natural, a mineração de texto também se baseia em avanços feitos em outras disciplinas da informática relacionadas, como o processamento de linguagem natural (PLN), como também explora técnicas e métodos das áreas de recuperação da informação, extração de informações e linguística computacional baseada em *corpus*. Um elemento-chave da mineração de texto é seu foco na coleção de documentos (METE *et al.*, 2010).

2.2.1 Coleção de documentos

Uma coleção de documentos pode ser qualquer agrupamento de documentos baseados em texto. Na prática, a maioria das soluções de mineração de texto é destinada a descobrir padrões em coleções de documentos.

No trabalho realizado por Kloptchenko *et al.* (2004) foram utilizadas técnicas de mineração de texto para estudar padrões ocultos sobre o desempenho financeiro futuro de empresas por meio de seus relatórios financeiros. Nesse cenário esses relatórios foram considerados uma coleção de documentos. Do ponto de vista dos autores, os relatórios anuais são um meio importante para a comunicação da empresa com suas partes interessadas e importante fonte de informações.

O número de documentos em tais coleções pode variar da casa dos milhares à casa dos milhões. Coleções de documentos podem ser estáticas, caso em que o complemento inicial de documentos permanece inalterado ou dinâmico, que é um termo aplicado às coleções de documentos caracterizadas pela inclusão de documentos novos ou atualizados ao longo do tempo. Coleções de documentos extremamente grandes bem como coleções de documentos com taxas muito altas de mudança de documento podem representar desafios de otimização de desempenho para vários componentes de um sistema de mineração de texto (FELDMAN; SANGER, 2006).

2.2.2 Etapas da mineração textos

Os sistemas de mineração de texto no geral são divididos em cinco etapas principais, sendo elas:

a) Extração de documentos

Os processos automatizados de classificação de documentos em que existe a necessidade de grande quantidade de dados para treinamento torna a etapa de extração e coleta de documentos importante no processo, como no trabalho de Ichinose e Shimada (2016). Esses autores propõem um método que utiliza informações extraídas de notícias para analisar o mercado de ações, tendo extraído 44.164 artigos de notícias de forma automática do *site Yahoo Japan Finance*.

b) Pré-processamento

A etapa do pré-processamento trata dos procedimentos necessários para organizar os dados para a próxima etapa da mineração de texto. A finalidade dessa etapa é formatar a informação original de maneira a torná-la compreensível aos métodos de mineração.

c) Tarefas de mineração

As tarefas básicas de mineração de texto consistem em mecanismos para a descoberta de padrões dentro de determinada coleção de documentos. Os três tipos mais comuns de padrões encontrados na mineração de texto são distribuições, conjuntos frequentes e quase frequentes e associações (TESO *et al.,* 2018).

d) Camada de apresentação

A camada de apresentação fornece funcionalidades de busca e pesquisa, além de ferramentas de visualização voltadas para o usuário.

e) Pós-processamento

A etapa de pós-processamento combina técnicas de refinamento que englobam métodos de filtragem de informações com redundância e agrupamento de dados associados. Essa etapa pode crescer em um sistema de mineração e com isso conceber um conjunto de aproximações para ordenação, poda, generalização, entre outras tarefas, levando à otimização dos sistemas que desenvolvem essa camada (ESKICI; KOÇAK, 2018).

2.2.3 Recuperação de informação

A recuperação de informação (RI) é uma área abrangente que se concentra notadamente em prover aos usuários o acesso simplificado às informações de seu interesse. A recuperação de informação aborda desde a representação, armazenamento, organização e acesso a itens de informação, a documentos, páginas *web*, catálogos *online*, registros estruturados e semiestruturados, objetos multimídia, etc. A representação e o arranjo dos itens de informação precisam prover os usuários de facilidade de acesso às informações de seu interesse (BAEZA-YATES; RIBEIRO-NETO, 2013). O aumento de conteúdo gerado pelos usuários na *web* na forma de notícias, comentários, *blogs*, redes sociais, *tweets*, fóruns, etc. resultou em um ambiente onde todos podem expressar publicamente sua opinião sobre eventos, produtos ou pessoas.

Essa riqueza de informações é potencialmente de vital importância para instituições e empresas, proporcionando-lhes maneiras de pesquisar seus

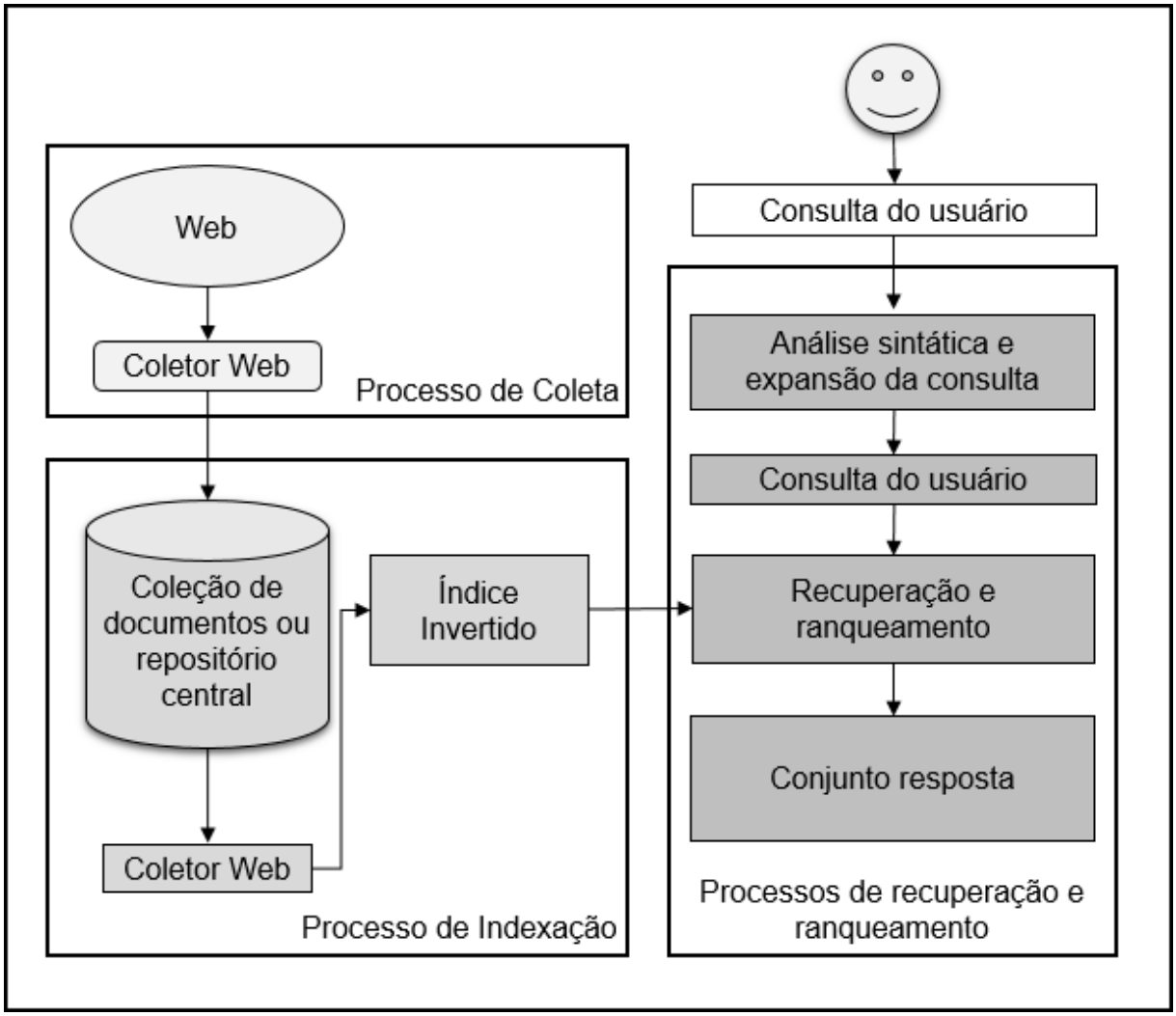
consumidores, gerenciar sua reputação e identificar novas oportunidades. Para muitas empresas, a opinião *online* se transformou em uma espécie de moeda virtual que pode fazer ou quebrar um produto no mercado (FELDMAN; SANGER, 2007). A análise de sentimentos, ou mineração de opinião, como também é conhecida, fornece mecanismos e técnicas por meio dos quais essa grande quantidade de informação pode ser processada e aproveitada.

A procura por informações é realizada há muitos anos e a ideia principal incide em encontrar documentos por meio de termos explicitados pelo usuário. O homem organiza informações para busca e emprego futuro. Um exemplo característico pode ser visto nos trabalhos efetivados por bibliotecários, assistentes e pesquisadores. Nesse aspecto também se pode citar o índice de um livro (CHOWDHURY, 2005). Com o aumento das bases de informações, surgiu a necessidade de implementação da primeira estrutura de acesso às informações armazenadas. Por séculos, esses sistemas foram criados e conduzidos manualmente, a partir de camadas de classificação tal como o índice alfabético. Contudo, com as melhorias tecnológicas e as mudanças nas necessidades individuais, fez-se necessária a melhora dessa área para que ela fosse capaz de atender de forma mais hábil às demandas de cada um. Com isso, na década de 1990 apareceram os sistemas de busca *online,* que permitiram mais facilidade no acesso à informação por parte de qualquer indivíduo com acesso à internet. Foi com o nascimento e a popularização da internet que a recuperação de informação ficou realmente popular. Nesse período apareceram os primeiros sistemas de busca *online*, como o *Google*.

Nesses sistemas é praticado o método mais conhecido de extração de informações, onde um termo é empregado para pesquisar um grande conjunto de documentos armazenados em uma base de dados. Na maior parte dos casos, a precisão nos resultados não é satisfatória, competindo ao usuário filtrar a informação e abstrair aquilo que é relevante ao que ele busca.

Baeza-Yates e Ribeiro-Neto (2013) asseguram que a recuperação de informações relevantes está diretamente relacionada à tarefa do usuário e com a forma de representação dos documentos utilizada no sistema de recuperação de informações. Segue na FIG. 2 a arquitetura em alto nível de um sistema de recuperação da informação.

Figura 2 - Arquitetura em alto nível de um sistema de recuperação da informação



Fonte: adaptada de Baeza-Yates e Ribeiro-Neto (2013).

O usuário de um sistema de recuperação de informações deve saber dizer ao sistema a informação exata que ele precisa. Em outras palavras, ele deve ser capaz de traduzir a sua necessidade em uma linguagem específica do sistema. Comumente, essa linguagem consiste em um determinado conjunto de palavras que semanticamente exprimem a necessidade do usuário (NASUKAWA; NAGANO, 2001). Já a representação lógica de documentos devido a razões históricas ocorre muito frequentemente a partir dos conjuntos de palavras-chave (que podem ser extraídas diretamente dos documentos analisados ou especificadas pelo usuário) e índices.

2.2.4 Processamento de linguagem natural

Processamento de Linguagem Natural (PLN) é a área de estudo que pesquisa como os computadores podem ser usados para entender e manipular textos ou linguagem natural para realizar tarefas úteis. Os pesquisadores do PNL buscam reunir conhecimentos sobre como os seres humanos entendem e usam a linguagem para que ferramentas e técnicas adequadas possam ser desenvolvidas de modo que os sistemas de computador compreendam e manipulem linguagens naturais para executar as tarefas desejadas. A fundação do PLN está em várias disciplinas, como, por exemplo, a Informática, Ciências da Informação, Linguística, Matemática, Engenharia Elétrica e Eletrônica, Inteligência Artificial, Robótica e Psicologia. Aplicações de PLN incluem vários campos de estudo, tais como tradução automática, processamento de texto em linguagem natural e sumarização, interfaces de usuário, recuperação de informações em vários idiomas, reconhecimento de fala, inteligência artificial e sistemas especialistas (CHOWDHURY, 2005).

No trabalho realizado por Abdullah, Rahaman e Rahman (2013) foram utilizadas técnicas de processamento de linguagem natural para extrair informações de textos de notícias coletadas da internet com o objetivo de identificar o impacto das notícias no mercado de ações e com isso identificar a tendência dessas ações. Para a realização dessa tarefa utilizou-se o *Apache OpenNLP*, que é um *kit* de ferramentas de aprendizagem de máquinas baseado em Java para processamento de linguagem natural.

Existem técnicas de processamento de linguagem natural que utilizam e produzem recursos linguísticos independentes do domínio, como tokenização, marcação de *tags* e análises sintáticas. Os documentos podem ser divididos em capítulos, seções, parágrafos, frases, palavras e até mesmo sílabas ou fonemas. A abordagem mais encontrada nos sistemas de mineração de texto envolve a quebra do texto em frases e palavras, que é chamado de tokenização. É comum que o processo de tokenização também extraia características (BATES, 1995).

A marcação de *tags* é a anotação de palavras com as *tags* adequadas com base no contexto em que elas aparecem. As *tags* dividem palavras em categorias com base no papel que desempenham na frase em que aparecem. As *tags* fornecem informações sobre o conteúdo semântico de uma palavra. Os substantivos

geralmente denotam coisas tangíveis e intangíveis, enquanto as preposições expressam relações entre coisas. A maioria dos conjuntos de *tags* usa as mesmas categorias básicas. O conjunto mais comum de *tags* contém “artigo, núcleo, verbo, adjetivo, preposição, número e substantivo”. E por fim a análise sintática realiza uma análise completa das frases de acordo com uma teoria gramatical do idioma em que se pretende trabalhar (FELDMAN; SANGER, 2006).

2.3 Aprendizado de máquina

Aprender é um feito multifacetado. Os métodos de aprendizagem contêm a obtenção de novos conhecimentos declarativos, o incremento de habilidades motoras e cognitivas por meio da instrução ou prática, o arranjo de novos conhecimentos em aspectos gerais e efetivas e a descoberta de novos fatos e teorias com base na observação e experimentação. Desde o início da era do computador, os pesquisadores têm se empenhado em inserir tais competências em computadores. Deliberar esse problema tem sido e permanece um objetivo de longo prazo muito desafiador e encantador no campo da inteligência artificial (IA) (SEBASTIANI, 2002). O estudo e a modelagem computacional dos processos de aprendizagem em suas múltiplas revelações estabelecem o objeto da aprendizagem de máquina.

O campo da aprendizagem de máquina é organizado em torno de três focos de pesquisa primária:

a) Estudos orientados a tarefas - o desenvolvimento e análise de sistemas de aprendizagem para melhorar o desempenho em um conjunto predeterminado de tarefas;
b) simulação cognitiva - a investigação e simulação computacional de processos de aprendizagem humana;
c) análise teórica - a exploração teórica do espaço de possíveis métodos de aprendizagem e algoritmos independentes do domínio da aplicação.

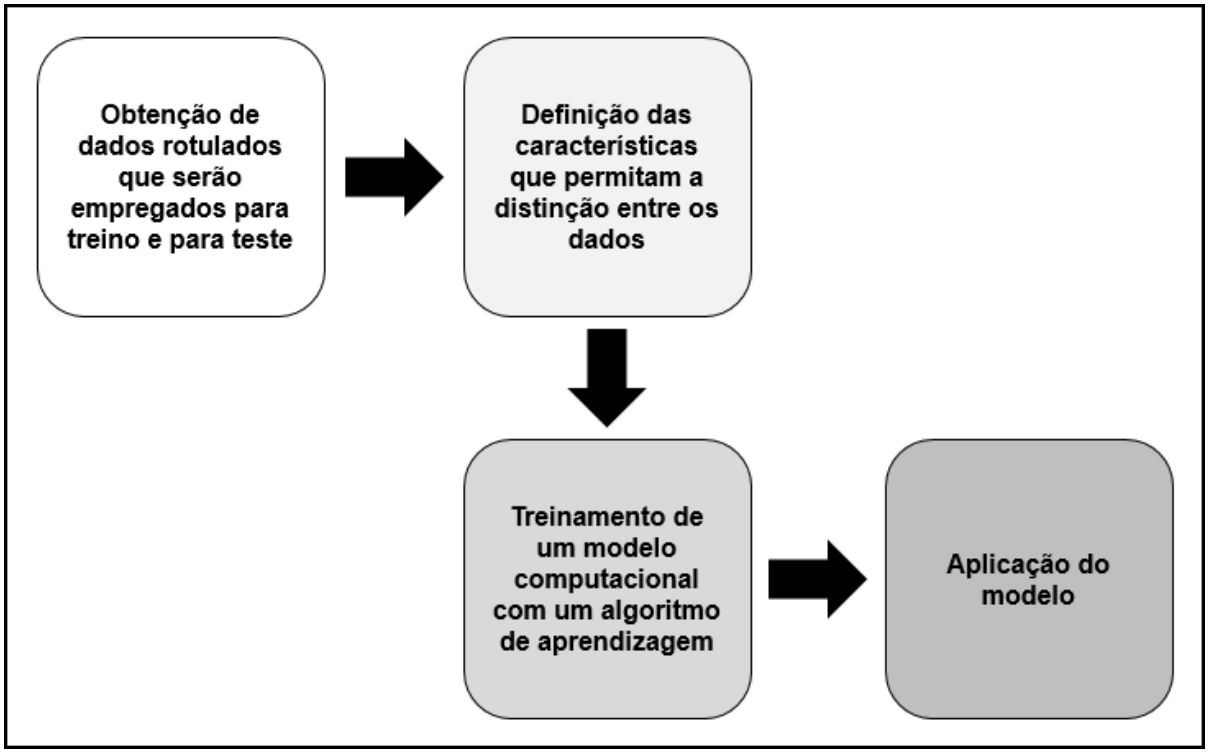
Embora muitos esforços de pesquisa se ocupem principalmente de um desses objetivos, o progresso em direção a um objetivo geralmente leva ao

progresso em direção a outro. Por exemplo, para investigar o espaço de possíveis métodos de aprendizagem, um ponto de partida razoável pode ser considerar o único exemplo conhecido de comportamento de aprendizado robusto, nomeadamente humano. Da mesma forma, as investigações psicológicas da aprendizagem humana podem ser auxiliadas por análises teóricas que podem sugerir vários modelos de aprendizagem plausíveis. A necessidade de adquirir determinada forma de conhecimento em algum estudo orientado a tarefas pode gerar novas análises teóricas ou estabelecer a questão: "como os humanos adquirem essa habilidade específica (ou conhecimento)"? Essa tricotomia de objetivos desafiadores e de apoio mútuo é um reflexo de todo o campo da inteligência artificial, cujas pesquisas de sistemas experientes, simulação cognitiva e estudos teóricos fornecem um ambiente de problemas e ideias (CARBONELL; MICHALSKI; MITCHELL, 1983).

2.3.1 Aprendizado supervisionado

A tarefa de identificar objetos e atribuí-los a uma classe conhecida *a priori* por um conjunto de exemplos rotulados é definido como aprendizado supervisionado ou simplesmente classificação. A classe, ou seja, o rótulo, determina o significado do objeto. Métodos supervisionados exigem uma quantidade de exemplos rotulados para treinar um classificador e gerar um modelo capaz de prever as classes de um conjunto de dados, em que as classes são desconhecidas. Contudo, adquirir exemplos de treinamento pode tornar-se um processo inviável e, além disso, a aprendizagem é cara e demorada se o conjunto de treino contém volumes enormes de dados. Alternativa é empregar métodos que exploram ambos os dados não rotulados e rotulados para classificação. As técnicas supervisionadas utilizam o termo supervisionado precisamente pelo fato de demandar uma fase de treinamento de um modelo com amostras previamente classificadas (TABOADA *et al.,* 2011). A metodologia para realizar a aprendizagem de máquina envolve quatro passos principais, que são apresentados na FIG. 3.

Figura 3 - Metodologia para realizar a aprendizagem de máquina



Fonte: autor.

2.3.2 Aprendizado não supervisionado

As técnicas não supervisionadas, ao contrário das supervisionadas, não necessitam de sentenças rotuladas previamente e treinamentos para a concepção de um modelo. Essa é uma das suas fundamentais vantagens, uma vez que dessa forma não deixa que a aplicação fique limitada ao conjunto para o qual foram treinados. Entre as técnicas não supervisionadas sobressaem aquelas com abordagens léxicas, baseadas em um dicionário léxico de sentimento, uma espécie de dicionário de palavras que, ao contrário de ter como conteúdo o significado de cada palavra, possui em seu lugar um sentido quantitativo (que pode ser um número entre -1 e 1, sendo -1 o valor sentimental mais negativo e 1 o valor mais positivo) ou mesmo valor qualitativo (positivo/negativo, feliz/triste). Abordagens léxicas ponderam que palavras individuais têm o que é chamado de polaridade prévia, que é uma orientação semântica independente de contexto e que pode ser expressada com um valor numérico ou classe (TABOADA *et al.*, 2011).

Para extrair com mais precisão opiniões e sentimentos de um texto, é muito importante construir um léxico de sentimento robusto. Dicionários de palavras que

expressam sentimentos são mais eficazes quando características específicas do domínio são levadas em consideração (KIM; JEONG; GHANI, 2014). A orientação semântica é uma medida da subjetividade e da opinião no texto que geralmente captura um fator avaliativo (positivo ou negativo) e potência ou força (grau em que a palavra, frase, oração ou documento em questão é positiva ou negativa) em relação a um assunto, uma pessoa ou uma ideia.

2.4 Robôs operando no mercado financeiro

Segundo Nigro (2018), as atividades com processos repetitivos e constantes são passiveis de automatização por meio de algoritmos, de tal maneira que esse é o motivo da utilização dos robôs em diversas áreas, automatizar ações que são repetitivas e constantes. No mercado financeiro não é diferente, a análise de gráficos, estatísticas, comparações de projeções passadas são alguns exemplos de processos a que profissionais do mercado financeiro constantemente prestam atenção. O problema é que os seres humanos são indivíduos sensíveis e tomam decisões muitas vezes com base em emoções. Com isso, Suhadolnik, Galimberti e Da Silva (2010) mostram como a inserção de "robôs" pode auxiliar na operação do mercado financeiro.

Fein (2015) acredita que a revolução tecnológica transformou o mercado de produtos e serviços de investimento de forma significativa. Novas ferramentas que utilizam a internet vieram possibilitar que investidores individuais comprem e vendam títulos diretamente no mercado de ações sem o conselho de um corretor, consultor de investimento ou outro intermediário. Muitos investidores individuais adquiriram a experiência e a autoconfiança para conduzir seus próprios programas de investimentos *online*. As decisões tomadas no mercado financeiro são muitas vezes influenciadas pelas emoções e sentimento dos investidores, portanto, investidores e pesquisadores têm como objetivo desenvolver modelos sistemáticos para reduzir o impacto das emoções nas decisões. Com isso o uso de sistemas algorítmicos torna-se alternativa viável (ARPACI; KARAOGLU; AYVAZ, 2017).

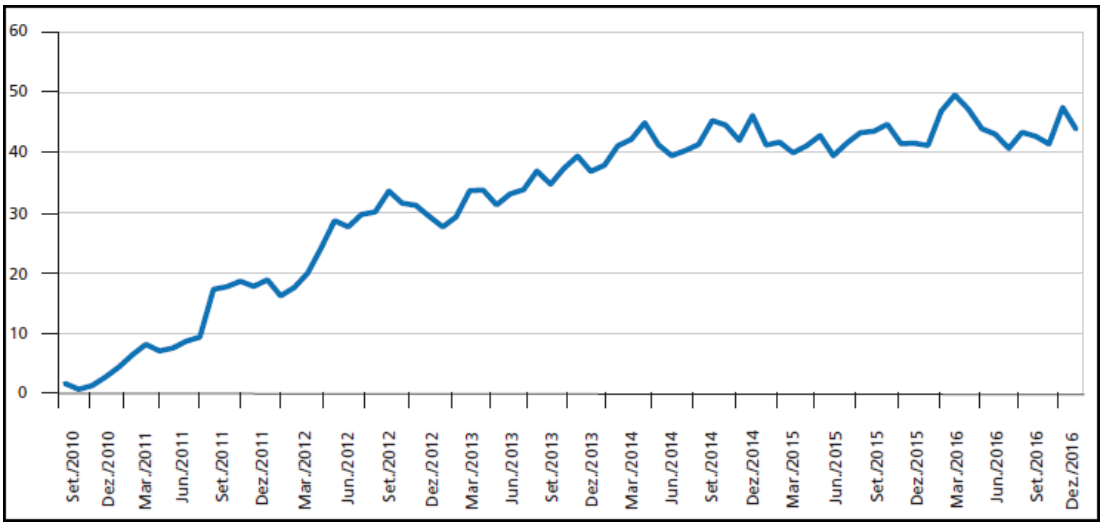
O tempo fracionado em milissegundos pode denotar grandes resultados financeiros. Dessa maneira, empreendendo a crescente flexibilização e liberação dos mercados financeiros em esfera global, a evolução tecnológica acabou se

impondo como uma das principais fronteiras da corrida entre os investidores visando ganhos cada vez mais elevados no menor intervalo de tempo possível. O movimento desse avanço acabou por causar um cenário onde negociações de alta frequência sejam empregadas como ferramentas de especulação e arbitragem entre diferentes ativos nos mercados, inflando o rendimento dos investidores que melhor dispõem dessas tecnologias. Portnoy (2011) define a estratégia de negociações de alta frequência como sendo algoritmos complexos desenvolvidos para transpor incontáveis probabilidades no mercado por meio dos computadores. Sua eficácia e eficiência se devem a antecipar o passo seguinte de um dado ativo financeiro nas próximas frações de segundo baseando-se em análises estatísticas das últimas *performances* e da atual condição do mercado.

De acordo com Paraná (2017), a utilização de robôs nos mercados brasileiros tornou-se possível de fato após 2010, que foi marcado pela implementação da modalidade de negociações *Co-location* nos segmentos BM&FBovespa, que ampliaram especialmente a capacidade de grandes investidores institucionais nacionais e estrangeiros para operar por meio da utilização de robôs. O *Co-location* é um serviço que aceita que os investidores instalem unidades de seus servidores dentro do próprio prédio da bolsa de valores, reduzindo a distância entre a ordem de negociação e sua execução no mercado. Esses algoritmos e dispositivos de transação automatizada, também conhecidos como robôs investidores, são responsáveis por mais de 40% de tudo que é comprado e vendido diariamente na bolsa de valores brasileira. No mercado americano, onde os investimentos em cabos de fibra ótica próprios e conexão extremamente rápidos por meio de micro-ondas são bilionários, consegue-se economizar dois a três milissegundos. Estima-se que esse percentual extrapole a marca dos 50%, tendo chegado ao seu apogeu por volta de 60-70% entre 2009 e 2014. Nos mercados europeus a média estimada de uso está em torno de 40% do total das negociações.

No GRÁF. 1 tem-se o número de negócios realizados na BM&FBovespa por meio de robôs investidores.

Gráfico 1 - Número de negócios realizados na BM&FBovespa por meio de robôs investidores em %



Fonte: Paraná (2017).

Um exemplo desse tipo de algoritmo é descrito por Ehrentreich (2006), Branch e Evans (2011). Trata-se do *Santa Fe Institute Artificial Stock Market* (SFI-ASM), que está sendo trabalhado desde a década de 1980. No SFI-ASM os investidores utilizam algoritmos para determinar quais estratégias adotarão e quais serão rejeitadas. O número de investidores no modelo é elevado, sendo que um modelo pode conter 10.000 agentes simulando o papel dos investidores.

Em relação à utilização de robôs, existem, por outro lado, estudos como os de Kirilenko e Lo (2013) e Ye, Yao e Gai (2012), que sustentam que o uso sistemático de robôs traz impactos negativos e que o uso de sistemas de negociações automatizadas nos mercados é predatório. Sustentam que as negociações de alta frequência colaboraram para a acentuada queda do mercado que eliminou inesperadamente quase US$ 1 trilhão das bolsas americanas em maio de 2010. E demonstram que negociações mais rápidas não acarretam necessariamente ganhos diretos ou em maiores volumes, mas aumentam a volatilidade nos mercados (PARANÁ, 2017).

Para mitigar esse tipo de ação e como medida de autorregulação e controle no mercado de capitais brasileiro a BM&FBovespa estabelece que os investidores que utilizem algoritmos computacionais forneçam aos reguladores internos e/ou externos a descrição de suas estratégias, com detalhamento de seus parâmetros de

negociações e limites, os principais mecanismos de controle de riscos e detalhes a respeito de como esses sistemas funcionam e como são testados.

2.5 Computação em nuvem e APIs para identificação de sentimentos

A computação em nuvem refere-se aos aplicativos fornecidos como serviços pela internet e ao *software* de *hardware* e sistemas nos centros de dados que fornecem esses serviços. Os serviços são referidos como *software* como serviço (SaaS), infraestrutura como serviço (IaaS) e plataforma como serviço (PaaS). Quando uma nuvem é disponibilizada de maneira pré-paga para o público em geral, é chamada de nuvem pública. Usa-se o termo nuvem privada para se referir a centros de dados internos de uma empresa ou outra organização, não disponibilizados para o público em geral (ARMBRUST *et al.,* 2010).

Segundo Srinivasan (2014), a computação em nuvem tornou-se popular por causa dos serviços confiáveis fornecidos por grandes empresas como *Amazon*, *Google*, *International Business Machines* (IBM) e *Microsoft*. Para fornecer um serviço de computação em nuvem confiável, o provedor deve investir grandes somas em infraestrutura, em que sua arquitetura inclui várias redundâncias. Cada um desses provedores de serviços globais possui vários *datacenters* espalhados por todo o mundo. Esses *datacenters* ajudam os consumidores da nuvem a atender aos requisitos governamentais de que os dados da nuvem devem residir dentro do país ou na região. Os *datacenters* distribuídos facilitam a redundância de armazenamento e ajudam na resposta de baixa latência.

Além dos grandes provedores como *Amazon*, *Google*, IBM e *Microsoft* de serviços de nuvem, há também vários provedores de serviços de nuvem de nicho, como *Salesforce*, *Apple*, *VMware*, *Dropbox* e *SoftLayer*. Essas empresas concentram-se em serviços específicos, como gerenciamento de relações com o cliente (CRM), distribuição de música, virtualização, armazenamento e servidores. Outro conjunto de empresas concentra-se no fornecimento de coordenação de serviços de terceiros para empresas que precisam de serviços em nuvem. Com a computação em nuvem as APIs da *web* tornaram-se reais, possibilitando implantar uma infraestrutura de forma global e ágil (BOTTA *et al.,* 2016).

2.5.1 Application programming interface

Uma interface de programação de aplicativo mais conhecido do inglês *application programming interface* (API) é uma especificação destinada a ser usada como uma interface por componentes de *software* para comunicação entre eles. Uma API pode incluir especificações para rotinas, estruturas de dados, classes de objetos e variáveis. Existem diferentes tipos de APIs, para sistemas operacionais, aplicativos ou *sites*. O *Windows*, por exemplo, tem muitos conjuntos de APIs usados por *hardware* e aplicativos do sistema. Quando um usuário copia e cola texto de um aplicativo para outro, por exemplo, é uma API que permite que isso funcione. A maioria dos sistemas operacionais, como o MS-*Windows*, fornece APIs, permitindo que os programadores escrevam aplicativos compatíveis com o ambiente operacional. Atualmente as APIs também são especificadas e disponibilizadas por *sites*. As APIs da *Amazon,* por exemplo, permitem que desenvolvedores usem a infraestrutura de varejo existente para criar lojas da na *web* de forma especializada. Desenvolvedores de *softwares* de terceiros também usam APIs expostas na *web* para criar soluções de *software* para usuários finais. Para diminuir o acoplamento entre as aplicações a tendência é que as APIs sejam disponibilizadas por meio de serviços REST API (ONG *et al.,* 2015).

2.5.2 REST API

REST é um acrônimo do inglês *Representational State Transfer*. É um estilo de arquitetura de *software* para aplicativos baseados em rede. Foi definido por Roy T. Fielding quando estava analisando as propriedades da *World Wide Web* e outras arquiteturas de rede e derivando as restrições arquitetônicas que tornaram o *world wide web* (WWW) bem-sucedido (FIELDING; TAYLOR, 2002). A arquitetura REST pode ser brevemente resumida como orientada a recursos em que cada entidade de informação importante ou coleção de entidades é considerada um recurso e é nomeada e endereçável, ou seja, seu conteúdo pode ser recuperado por seu endereço. Por exemplo, os correios podem definir o número de um código de endereçamento postal (CEP) como recurso e os clientes podem acessar esse recurso pelo endereço http://viacep.com.br/ws/30160000/json. A parte do endereço

http://viacep.com.br/ é a parte fixa, que é definido como o *endpoint*; e a parte variável /ws/30160000/json definida como recurso.

Um recurso pode retornar seu conteúdo em diferentes formatos. Utilizando o endereço http://viacep.com/ws/30160000/json pode-se definir que o retorno da REST API seja no formato *JavaScript Object Notation* (JSON). O JSON é em formato de texto e completamente independente de linguagem de programação, pois usa convenções que são familiares às linguagens C e familiares e está constituído em duas estruturas, uma coleção de pares nome/valor e uma lista ordenada de valores (ECMA, 2017). Segue na FIG. 4 um exemplo do retorno de uma REST API no formato JSON.

Figura 4 - Exemplo de retorno de uma REST API no formato JSON


```
{
    "cep": "30160-000",
    "logradouro": "Praça Rui Barbosa",
    "complemento": "",
    "bairro": "Centro",
    "localidade": "Belo Horizonte",
    "uf": "MG",
    "unidade": "",
    "ibge": "3106200",
    "gia": ""
}
```


Fonte: autor.

Alterando a parte referente ao recurso do REST API no endereço http://viacep.com.br/ws/30160000/xml pode-se definir que o retorno da REST API passe a ser no formato *Extensible Markup Language* (XML). O XML é um formato de texto simples e muito flexível, originalmente projetado para enfrentar os desafios da publicação eletrônica em grande escala (LIAM, 2018). Segue na FIG. 5 um exemplo do retorno de uma REST API no formato XML.

Figura 5 - Exemplo de retorno de uma REST API no formato XML

```xml
<xmlcep>
   <cep>30160-000</cep>
   <logradouro>Praça Rui Barbosa</logradouro>
   <complemento/>
   <bairro>Centro</bairro>
   <localidade>Belo Horizonte</localidade>
   <uf>MG</uf>
   <unidade/>
   <ibge>3106200</ibge>
   <gia/>
</xmlcep>
```

Fonte: autor.

O *design* da API REST inclui uma especificação dos formatos de representação permitidos para cada par recurso/operação. Os serviços da *web* de estilo REST tornaram-se alternativa popular aos serviços baseados em SOAP e são considerados mais leves e fáceis de usar (JELIAZKOVA; JELIAZKOV; AMBIT REST, 2011).

Seguem-se alguns exemplos populares de REST APIs:

O *site ProgrammableWeb* (https://www.programmableweb.com/) monitora mais de 19.593 APIs e possui uma classificação das APIs mais populares. Entre elas estão:

1º - **APIs do *Google Maps*:** permitem que os desenvolvedores incorporem o *Google Maps* em páginas da *web* usando uma interface *JavaScript*.
Endereço: https://developers.google.com/maps/
Categoria: mapeamento.

2º - **APIs do *Twitter*:** permitem que os desenvolvedores acessem os principais dados do *Twitter* e a API de pesquisa fornece métodos para os desenvolvedores interagirem com os dados de pesquisa e tendências do *Twitter*.
Endereço: https://dev.twitter.com/rest/public
Categoria: social.

3º - **APIs do *YouTube*:** permitem que os desenvolvedores integrem vídeos e funcionalidades do *YouTube* em *sites* ou aplicativos.
Endereço: https://developers.google.com/youtube/
Categoria: vídeo.

4º. **APIs do *Flickr*:** as APIs do *Flickr* são usadas pelos desenvolvedores para acessar os dados da comunidade de compartilhamento de fotos do *Flick*.
Endereço: http://www.flickr.com/services/api/
Categoria: fotos.

5º - **APIs do *Facebook*:** permitem que os desenvolvedores conectem interfaces e desenvolvam em várias plataformas.
Endereço: https://developers.facebook.com/
Categoria: social.

2.5.3 APIs para identificação de sentimentos

Com a popularização da computação em nuvem os provedores de serviços em nuvem passaram a oferecer diferentes tipos de serviços por meio de APIs que são disponibilizadas para os clientes. Neste tópico serão descritas APIs para identificação de sentimentos dos quatro maiores provedores de serviços em nuvens: *Amazon*, *Microsoft*, *Google* e IBM.

Amazon

Por meio da *Amazon Web Services*, provedora de serviços de nuvem da *Amazon* é disponibilizado API "*Amazon Comprehend*". Com essa API pode-se determinar se o sentimento é positivo, negativo, neutro ou misto. A abordagem da *Amazon* é a mais simples: fornece-se uma *string* de texto e o idioma para ser usado para análise. Atualmente, apenas inglês e espanhol são reconhecidos. Pode-se consultar a documentação sobre a utilização da API diretamente no *site* da *Amazon* no endereço https://docs.aws.amazon.com/comprehend/latest/dg/how-sentiment.html.

Microsof

A API da *Microsoft* é um conjunto de serviços da *web* de análise de texto criados com algoritmos de aprendizado de máquina da própria *Microsoft*. Sua API pode ser usada para analisar texto não estruturado para tarefas como análise de sentimento, extração de frase-chave e detecção de idioma. Nenhum dado de treinamento é necessário para usar essa API. Ela usa técnicas avançadas de processamento de linguagem natural para oferecer as melhores previsões da classe. A API retorna uma pontuação numérica entre zero e um. Pontuações próximas de um indicam sentimento positivo, enquanto pontuações próximas de zero indicam sentimento negativo. Pontuação de 0,5 indica a falta de sentimento. Pode-se consultar a documentação sobre a utilização de sua API diretamente no *site* da *Microsoft* no endereço HTTPS://docs.microsoft.com/en-us/azure/cognitive-services/text-analytucs/overview.

Google

A API para identificação de sentimentos do *Google* é conhecida como *Google Cloud Natural Language*. Ela aceita um texto simples, documento *HyperText Markup Language* (HTML) ou uma referência a um documento localizado no armazenamento do *Google Cloud* e o tipo de codificação do texto, que é importante em termos de cálculo de compensações. Pode-se especificar o idioma para ser usado na análise, por padrão, o que é detectado automaticamente. O algoritmo da *API* do *Google Cloud Natural Language* divide a variação do sentimento entre -1 e 1. Pontuações próximas de um indicam sentimento positivo, enquanto pontuações próximas de -1 indicam sentimento negativo. Pode-se consultar a página de documentação sobre a utilização da API no endereço https://cloud.google.com/natural-language/docs/?hl=pt-br.

IBM

IBM Watson Tone Analyzer faz parte da cadeia de ferramentas do *IBM Watson Developer Cloud,* que oferece serviços em nuvem, como serviços para processamento de linguagem natural, fala, visão e percepção de dados para o desenvolvimento de aplicativos cognitivos. Cada serviço do *Watson* fornece uma REST API para interagir com o serviço escolhido. Uma dessas APIs é chamada de

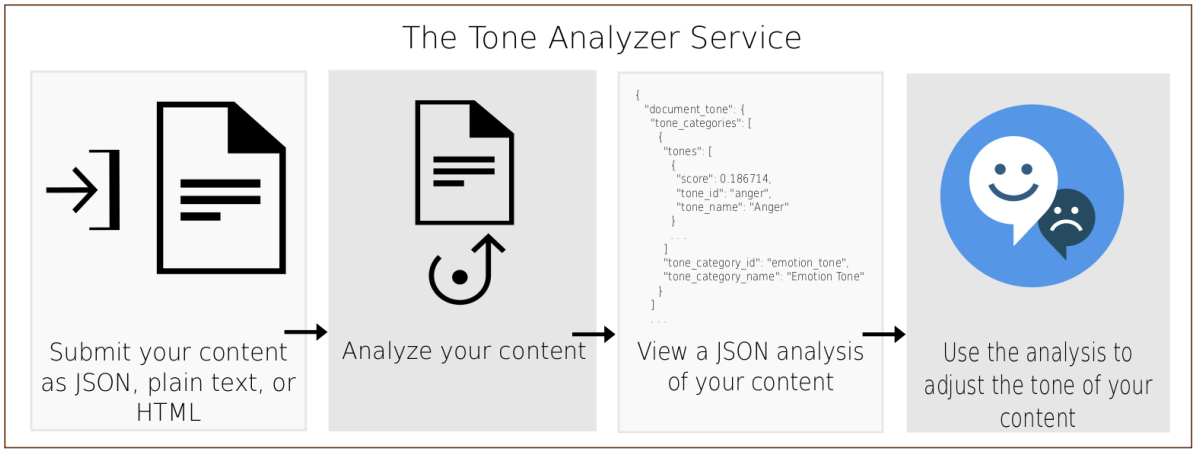
Tone Analyzer, que analisa o conteúdo de um texto, atribuindo valores diferentes às sentenças com o conjunto de emoções, que são alegria, tristeza, medo, raiva e desgosto. Cada emoção recebe um valor entre zero e um, sendo que valor menor que 0,5 indica que o sentimento provavelmente não está presente no texto; maior que 0,5 que provavelmente o sentimento está presente no texto; e valor maior que 0,75, que é muito provável que o sentimento esteja presente no texto. A IBM fornece detalhes técnicos sobre como essas pontuações são computadas em seu manual *online* (KAMINSKI, 2018).

O serviço IBM *Watson Tone Analyzer* é baseado na teoria da Psicolinguística, um campo de pesquisa que explora a relação entre o comportamento linguístico e teorias psicológicas. Para obter a pontuação da emoção de um texto a IBM usa um conjunto baseado em generalização empilhada. Esta usa um modelo de alto nível para combinar modelos de nível inferior para alcançar mais precisão da previsão. Características tais como n-gramas (unigramas, bigramas e trigramas), pontuação, *emoticons*, palavrões, cumprimentos (como "olá", "oi" e "obrigado") e polaridade são alimentadas em algoritmos de aprendizagem de máquina para classificar as categorias dos sentimentos (IBM, 2018a).

O *endpoint* do REST API do *IBM Watson Tone Analyzer* é definido como HTTPS://gateway.watsonplatform.net/tone-analyzer/api/. O recurso do REST API do IBM *Watson Tone Analyzer*, que é a parte variável, é composto pelos seguintes parâmetros: v3/tone?version=2017-09-21&text=Texto, o endereço completo composto pelo *Endpoint;* e recurso para acessar a REST API é definido como HTTPS://gateway.watsonplatform.net/tone-analyzer/api/v3/tone?version=2017-09-21&text=Texto.

Para extrair a emoção de um texto pode-se enviar na entrada do serviço REST API um JSON, texto simples ou HTML que contém o conteúdo para a extração do sentimento. O serviço aceita até 128 KB de texto e a utilização da API é gratuita para até 50 mil requisições por mês. O serviço analisa o conteúdo do texto e retorna o resultado em JSON com os índices dos sentimentos calculados para o texto passado na entrada. A FIG. 6 mostra um exemplo do processo para a extração do sentimento de um texto por meio da API do IBM *Watson Tone Analyzer.*

Figura 6 - Processamento do texto por meio da API do *IBM Watson Tone Analyzer*



Fonte: IBM (2018a).

Seguem-se alguns trabalhos que utilizaram a API do *IBM Watson Tone Analyzer.*

Agarwal e Sureka (2016) utilizaram a API do IBM *Watson Tone Analyzer* com o objetivo de identificar e classificar *posts* em redes sociais com conotações racistas e radicais. Demonstraram a eficácia dos traços de personalidade dos autores para identificar *posts* com tal intenção.

Mostafa *et al.* (2016) utilizaram o IBM *Watson Tone Analyzer* para identificar emoção nas interações textuais em um sistema *online*, com base em trabalhos anteriores nessa área que mostram forte correlação entre o que os usuários digitam e sua personalidade.

No trabalho de Johnson *et al.* (2017) foi utilizada a API do *Speech-To-Text* do IBM *Watson* para converter o áudio em texto e utilizar este texto como entrada para o IBM *Watson Tone Analyzer*, que por sua vez fornece o sentimento contido no texto. Dependendo do valor dado aos cinco possíveis sentimentos, ele é classificado como positivo ou negativo e as alterações de um avatar são influenciadas dependendo do tipo de interação com que o sistema classificou as palavras do usuário.

Michellc *et al.* (2017) realizou um experimento com textos publicados em redes sociais baseado em opiniões e sentimentos utilizando a API do *IBM Watson Tone Analyzer*, com o objetivo de estabelecer um modelo para geração de informações que vão desde índices de satisfação até as tendências da opinião pública.

3 TRABALHOS RELACIONADOS

Fan e Watanabe (2012) propuseram a aplicação de métodos de classificação e regressão para prever os preços das ações com base em artigos de notícias, experiências com dados reais e artigos de notícias. Hagenau, Liebmann e Neumann (2013) demonstraram que a combinação de métodos de extração de características aumenta a precisão da classificação. Melhor seleção de características melhora significativamente a precisão da classificação, pois a abordagem permite reduzir o número de características menos significativas, portanto, pode limitar os efeitos negativos ao aplicar as técnicas de aprendizado de máquina na classificação dos textos. Geva e Zahavi (2014) mostram que a integração de dados do mercado financeiro com dados textuais extraídos de *feeds* de notícias contribui para melhorar o desempenho da modelagem e que o uso de mais representações de dados textuais melhora a precisão preditiva.

Alostad e Davulcu (2015) propuseram um sistema para prever a tendência do preço das ações por hora com base na análise textual do conteúdo dos artigos de notícias que mencionam uma ação ou empresa dona dessa ação. Demonstraram que o uso de notícias mais recentes produz aumento estatisticamente significante da precisão da predição em comparação ao uso de todas as notícias indiscriminadamente. Gunduz e Cataltepe (2015) estudaram os efeitos dos artigos de notícias sobre a ação BIST100 da Bolsa de Valores de Istambul. Além das informações sobre os preços, as palavras que ocorreram nos artigos de notícias do dia anterior foram usadas como características. Duong, Nguyen e Dang (2016) comprovaram a correlação entre as notícias financeiras e os preços das ações no índice VN30 (índice do mercado de ações do Vietnã).

4 REVISÃO SISTEMÁTICA DA LITERATURA

4.1 Introdução

O conteúdo dos meios de comunicação *online* tem se mostrado um importante fator que molda o sentimento dos investidores por meio do sentimento que o texto das notícias carrega (MAO; COUNTS; BOLLEN, 2011). A economia comportamental afirma que as emoções podem afetar profundamente o comportamento e a tomada de decisão de um indivíduo (BOLLEN; MAO; ZENG, 2011; NOFSINGER, 2005). Se a emoção do indivíduo investidor pode afetar a forma como ele reage às novas informações, é aceitável que os sentimentos coletivos dos investidores possam influenciar a dinâmica do mercado de ações (OLIVEIRA; CORTEZ; AREAL, 2013). Os artigos de notícia são primeiro interpretados por investidores e traduzidos em sentimento do mercado em que os investidores, em seguida, tomam suas decisões com base no sentimento e interpretações. Com isso, os preços de mercado agregam as ações de cada investidor e refletem-nas nos movimentos do preço final (LI, X. *et al*., 2014).

Para melhor abrangência das técnicas indicadas na literatura aplicadas na identificação de tendências no mercado de ações por meio de notícias publicadas na internet realizou-se revisão sistemática de literatura. Esta foi organizada da seguinte forma: a seção 4.2 aborda a metodologia utilizada no desenvolvimento da revisão sistemática. A seção 4.3 apresenta os resultados. Na seção 4.4 faz-se a conclusão.

4.2 Metodologia da revisão sistemática da literatura

A revisão sistemática responde a uma pergunta claramente formulada, empregando métodos sistemáticos e explícitos para identificar, selecionar e avaliar criticamente estudos relevantes e colher e ponderar dados dos estudos abrangidos na revisão (KITCHENHAM, 2004). Segundo Brito (2016), existem motivos para a realização de uma revisão sistemática da literatura, sendo algumas delas:

a) Sintetizar as evidências existentes a propósito de uma tecnologia;

b) identificar brechas nas pesquisas existentes e indicar novos assuntos para debate;
c) prover um arcabouço a fim de posicionar novas atividades futuras.

O objetivo desta pesquisa foi investigar os estudos que relacionam o impacto das notícias sobre empresas e a oscilação do preço de suas ações no mercado de capitais que utilizam técnicas de mineração de texto e aprendizado de máquina.

4.2.1 Questões de pesquisa

Para o protocolo da revisão sistemática de literatura adotado, no QUADRO 1 propõem-se as seguintes questões de pesquisa para atender ao objetivo deste estudo:

Quadro 1 - Questões de pesquisa da revisão sistemática da literatura

ID	Questão
1	Quais são as técnicas de pré-processamento e representação de características mais utilizadas no tratamento de notícias publicadas na internet com o objetivo de identificar tendências no mercado de ações?
2	Quais são os classificadores mais utilizados para a identificação de tendências no mercado de ações utilizando-se de notícias publicadas na internet?
3	Qual a relação entre o tamanho do *corpus* e a precisão retornada por um classificador?

Fonte: autor.

4.2.2 Amplitude da pergunta

- **População**: trabalhos que atuem com a mineração de texto e processamento de linguagem natural, notícias publicadas na internet, recuperação da informação e aprendizado de máquina;
- **intervenção**: processos de mineração de texto em notícias publicadas na internet;
- **comparação**: documentar as diferentes técnicas utilizadas para identificar a relação entre o impacto das notícias publicadas na internet e o mercado de ações;

- **resultados**: Identificar as técnicas mais utilizadas na previsão de tendências no mercado de ações por meio de notícias publicadas na internet.

4.2.3 Planejamento da revisão sistemática

O procedimento adotado para o desenvolvimento da revisão sistemática foi realizado de acordo com o modelo de protocolo apresentado por Kitchenham (2004). Como estratégia de busca foi utilizada a busca automática em bases eletrônicas, tendo sido considerados os elementos descritos a seguir:

a) Seleção das fontes de pesquisa:
   A escolha do grupo de bases eletrônicas para as buscas é baseada em (DIESTE; GRIMÁN; JURISTO, 2009), que determina os seguintes critérios: i) disponibilidade dos estudos primários; ii) cobertura das publicações e conferências relevantes na área; iii) a busca por estudos em inglês por ser a língua amplamente adotada nos principais eventos e periódicos da área. Seguindo esses critérios, as fontes selecionadas foram:
   - ACM *Digital Library* (<http://dl.acm.org/>)
   - IEEE *Xplore* (<http://ieeexplore.ieee.org/Xplore/home.jsp >)
   - *Science Direct* (<http://www.sciencedirect.com/ >)

b) Palavras-chave:
   "*Text Mining*" *OR* "*Sentiment Analysis*", "*Stock Market*", "*news*".

c) *String* de busca:
   A *string* de busca relacionou as palavras-chave pelo operador lógico *AND* e, além disso, foram incluídos os possíveis sinônimos relacionados pelo operador lógico *OR*. A *string* resultante foi **(("*Text Mining*" *OR* "*Sentiment Analysis*") *AND* ("*Stock Market*") *AND* ("*news*")).**

4.2.4 Critérios para seleção

4.2.4.1 Critério de inclusão

a) Artigos que relacionavam notícias e seu impacto no mercado de ações, utilizando técnicas de mineração de texto e/ou analise de sentimentos.

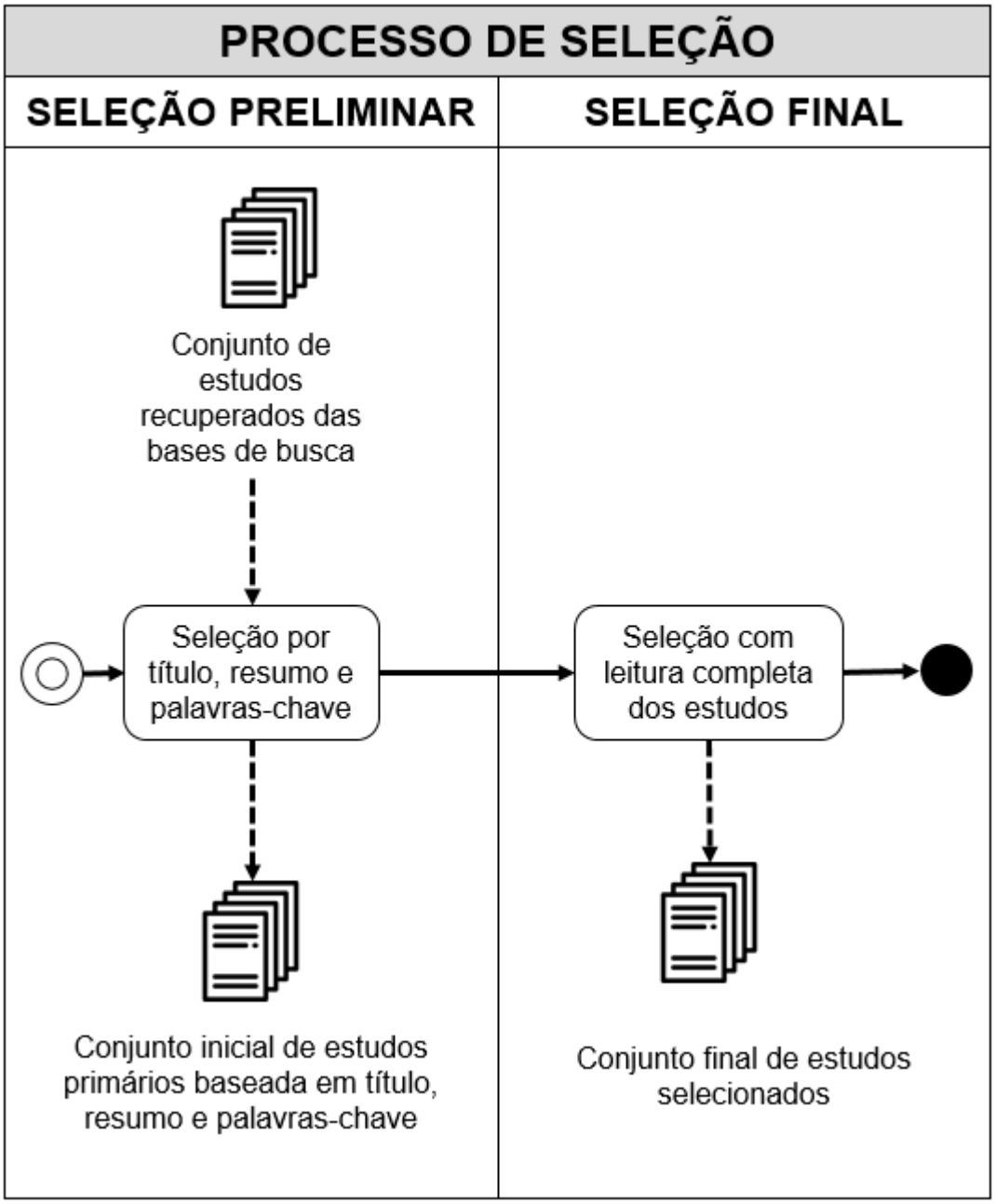
4.2.4.2 Critério de exclusão

a) Critério de exclusão EC1: artigos que não utilizassem notícias publicadas na internet como meio para identificação de tendências no mercado de ações e os que utilizassem *Twitter*, *Facebook*, *blogs* e demais mídias;
b) critério de exclusão EC2: o conteúdo do artigo precisava detalhar o uso de um algoritmo computacional;
c) critério de exclusão EC3: o artigo não era escrito em inglês/português;
d) critério de exclusão EC4: artigos com data de publicação inferior ao ano de 2012. Esse critério foi aplicado diretamente durante as buscas nas bases de dados, portanto, valeu para aqueles estudos que fossem retornados mesmo após filtro aplicado nas bases de dados;
e) critério de exclusão EC5: o texto completo do artigo não estivesse disponível.

4.2.4.3 Processo de seleção

A seleção dos estudos primários foi dividida em duas fases principais: o processo de seleção preliminar, que inclui a leitura do título, resumo e palavras-chave, e o processo de seleção final, que inclui a leitura do texto completo dos estudos. A FIG. 7 ilustra como as atividades envolvidas no processo de seleção foram organizadas.

Figura 7 - Processo de seleção dos artigos na revisão sistemática da literatura



Fonte: autor.

Uma vez recuperados os estudos primários das bases de pesquisa, foram eliminados os duplicados e passou-se para a atividade de seleção com base nos critérios de inclusão/exclusão em confronto com as informações disponíveis no título, resumo e palavras-chave. Para a inclusão e exclusão final dos estudos primários incluídos na seleção preliminar considerou-se o confronto dos critérios de inclusão/exclusão com o texto completo desses estudos.

4.2.4.4 Realização das buscas

A *string* de pesquisa foi adaptada para cada base e foi testada verificando se os estudos retornados em cada busca continham todos os termos esperados. A seguir são apresentadas as *strings* finais adaptadas para cada base utilizada, bem como comentários adicionais sobre particularidades e dificuldades de busca em cada base.

ACM Digital Libray

Detalhe: nessa base foi necessário colocar os operadores lógicos (*OR/AND*) em *lower case*. Além disso, a busca foi realizada na seção de busca avançada. A interface com os resultados permite filtrar o resultado a partir de um ano selecionado.

String final adaptada: "*Text Mining*" or "*Sentiment Analysis*" +"*Stock Market*" + *news.*

Filtro selecionado: artigos publicados a partir de 2012.

IEEEXplore

Detalhe: a *string* foi aplicada na seção de busca avançada com a opção *Command Search*" e seguiu as recomendações de estrutura presentes na base. A interface com os resultados permite filtrar o resultado a partir de um ano selecionado. *String* final adaptada: *(("Text Mining" OR "Sentiment Analysis") AND ("Stock Market") AND ("news")).*

Filtro selecionado: artigos publicados a partir de 2012.

Science Direct

Detalhe: a pesquisa foi realizada no modo *expert*. A interface com os resultados permite filtrar o resultado a partir de um ano selecionado. *String* final adaptada: *(("Text Mining" OR "Sentiment Analysis") AND ("Stock Market") AND ("news")).*

Filtro selecionado: artigos publicados a partir de 2012.

A busca foi realizada no dia 09/03/2017 às 09:00h da manhã, desconsiderando estudos anteriores ao ano de 2012. Foram recuperados ao todo 210 estudos primários para o processo de seleção (inclusão/exclusão) descrito no planejamento.

4.2.5 Estratégias de extração de dados e sumarização dos resultados

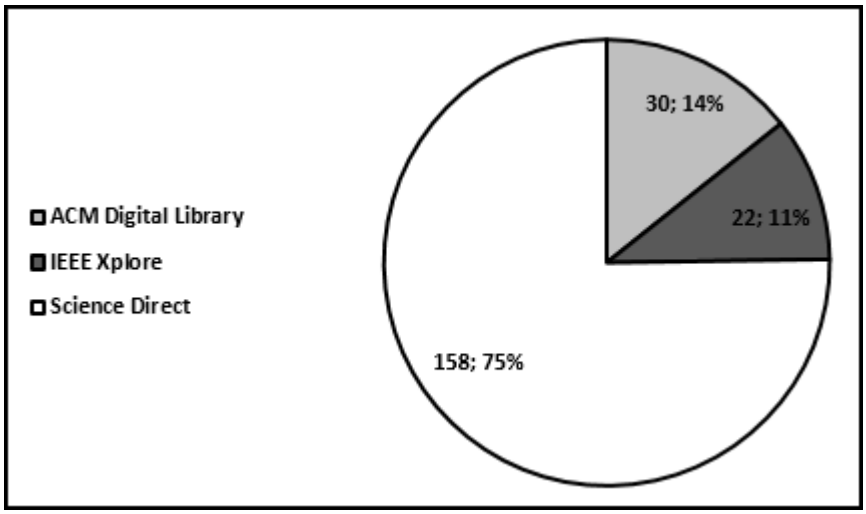
A estratégia de extração de evidências adotada foi a de formulários de extração de dados, nos quais durante a leitura dos textos completos o pesquisador alimenta com evidências encontradas nos estudos os campos do formulário. Para dar suporte à execução das estratégias, as seguintes ferramenta foram adotadas:

Mendeley: organização das referências;
Microsoft Excel: tabulação dos dados numéricos para geração de gráficos.

4.3 Resultados da revisão sistemática da literatura

Para o cumprimento da revisão sistemática, foi obedecido o protocolo instituído na seção 4.2. As pesquisas foram realizadas com os critérios de busca empregando também as parametrizações de pesquisa das bases de dados. Os resultados foram adquiridos na ordem de classificação por importância, de acordo com o próprio mecanismo de busca. A execução da *string* de busca nas bases escolhidas para a implementação desta pesquisa retornou o total de 210 trabalhos distribuídos entre os anos de 2012 e 2017 entre as bibliotecas digitais, conforme resumido no GRÁF. 2.

Gráfico 2 - Distribuição dos artigos encontrados na pesquisa automática por base de busca



Fonte: dados da pesquisa.

Com base nos critérios de inclusão e exclusão o filtro foi aplicado na ordem de leitura previamente definida: na seleção preliminar foi primeiramente utilizado o título dos trabalhos; em seguida, o resumo e palavras-chave, que resultaram em 63 estudos selecionados. E na seleção final foi realizada a leitura dos textos completos, reduzindo-se o *corpus* inicial da pesquisa para 29 estudos.

A seguir, na TAB. 1 são apresentados os resultados do processo de seleção.

Tabela 1 - Quantidade de artigos obtidos após aplicação dos critérios de inclusão

Base de Busca	Itens Retornados	Incluídos Seleção Preliminar	Incluídos Seleção Final
ACM Digital Library	30	14	5
IEEE Xplore	22	17	9
Science Direct	158	32	13
Total	**210**	**63**	**29**

Fonte: dados da pesquisa.

Na seleção preliminar foram incluídos 63 estudos e excluídos 147. Em sequência, na seleção final restaram 29 estudos. A TAB. 2 apresenta o grupo de estudos excluídos.

Tabela 2 - Artigos excluídos

Critério de Exclusão	Excluídos Seleção Preliminar	Excluídos Seleção Final	Total
EC1	147	34	175
EC2	0	0	0
EC3	0	0	0
EC4	0	0	0
EC5	0	0	0
TOTAL	**147**	**34**	**175**

Fonte: dados da pesquisa.

A qualidade da interpretação de sentimentos nas notícias publicadas na internet pode determinar a previsibilidade dos mercados financeiros. Com isso pesquisadores voltaram sua atenção para os diferentes aspectos desse problema. A FIG. 8 exibe mais interesse em estudos para identificação de tendências no mercado de ações por meio de notícias publicadas na internet entre os anos de 2014 e 2016.

Figura 8 - Artigos por ano

2012	2013	2014	2015	2016	2017
3	3	8	6	8	1

Fonte: autor.

Baseado nos trabalhos estudados, listaram-se as principais técnicas utilizadas na identificação de tendências no mercado de ações por meio de notícias publicadas na internet.

Todos os trabalhos utilizaram pelo menos duas fontes de dados como entrada, os dados de texto de notícias extraídas da internet e os dados do mercado de ações, ou seja, as cotações das ações. Em relação às notícias a maioria dos estudos utiliza notícias financeiras, pois consideram que estas provocam menos ruído em comparação às notícias gerais. São extraídos das notícias o texto principal do corpo da notícia ou o título também conhecido como manchetes. As manchetes das notícias são muitas vezes utilizadas por serem mais diretas e, portanto, têm menos ruído causado pelos textos detalhados (GEVA; ZAHAVI, 2014; KHADJEH NASSIRTOUSSI *et al.,* 2015; MUKUND *et al.*, 2016; TIREA; NEGRU, 2016).

A outra fonte de dados de entrada vem dos valores numéricos dos mercados financeiros, sob a forma de preços das ações ou índices. Esses dados são usados

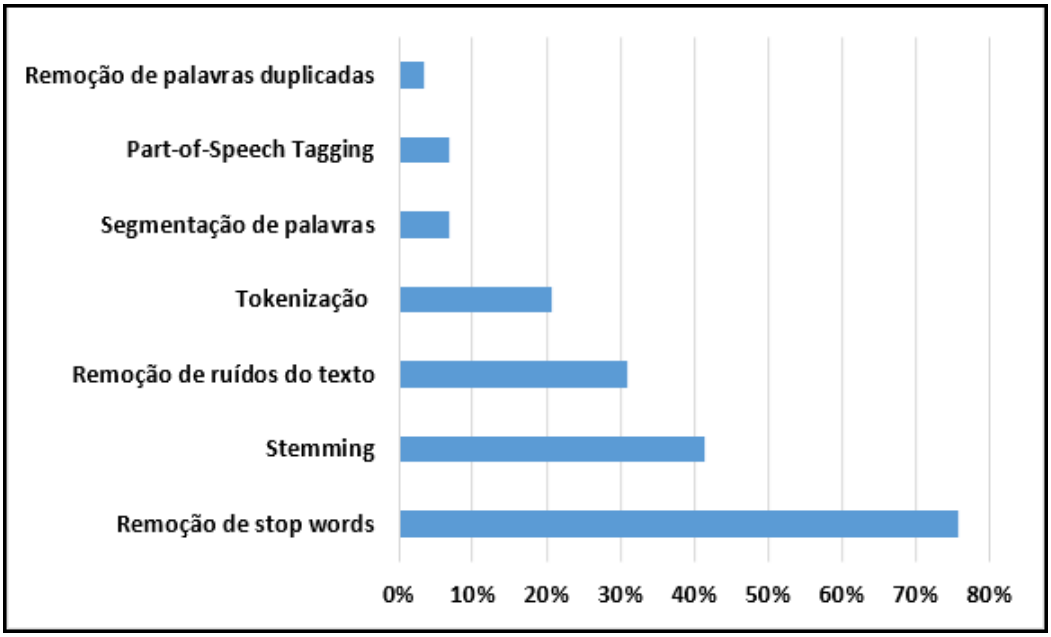
principalmente com o propósito de treinar os algoritmos de aprendizado da máquina e para fins de predição. Como resultado e análise da revisão sistemática, as três perguntas propostas anteriormente foram respondidas.

Questão 01 - Quais são as técnicas de pré-processamento e representação de características mais utilizadas no tratamento de notícias publicadas na internet com o objetivo de identificar tendências no mercado de ações?

Na mineração de textos a etapa do pré-processamento trata dos procedimentos necessários para organizar os dados para a próxima etapa da mineração. A finalidade dessa etapa é formatar a informação original de maneira a torná-la compreensível aos métodos de mineração. Uma vez que os dados de entrada estão disponíveis, eles devem ser preparados para que possam ser utilizados por determinado algoritmo. Isso para os dados de texto significa transformar o texto não estruturado em um formato representativo estruturado e que possa ser processado. Com isso, a fase de pré-processamento tem impacto significativo nos resultados.

Nos estudos analisados verificou-se que as técnicas de pré-processamento mais utilizadas foram a remoção de *stop words*, utilizada em 76% dos estudos; redução das palavras, utilizando técnicas de *stemming,* citada em 41% dos estudos; remoção de ruídos que geralmente incluem *tags html*, *links* e propagandas, que foi citada em 31% dos estudos; e a tokenização, que foi citada em 21% dos estudos, conforme se pode verificar no GRÁF. 3.

Gráfico 3 - Técnicas de pré-processamento mais utilizadas



Fonte: autor.

No pré-processamento, uma etapa importante é a seleção de características do texto, pois o número de palavras diferentes é grande mesmo em documentos relativamente pequenos, como em artigos de notícias curtas. Além disso, os vetores de representação de documentos, embora esparsos, ainda podem ter centenas e milhares de palavras. A maioria dessas palavras é irrelevante para a tarefa de categorização e pode ser descartada sem prejudicar a *performance* do classificador, podendo até resultar melhorias devido à redução de ruído (FELDMAN; SANGER, 2006; LI, X. *et al*. 2014).

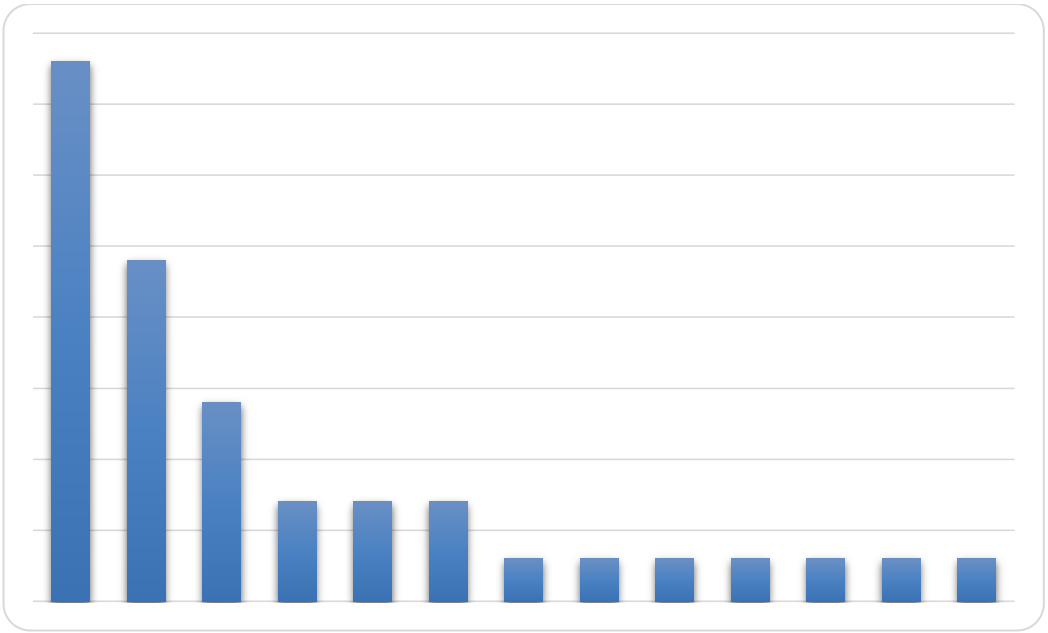
Depois de determinar o número mínimo de características, cada uma delas precisa ser representada por um valor numérico para que ele possa ser processado por algoritmos de aprendizado de máquina. Esse valor numérico atribuído funciona como uma pontuação ou um peso. Os métodos para dar pesos às características podem variar. Nos trabalhos de Fan e Watanabe (2012), Nizer e Nievola (2012), Hagenau, Liebmann e Neumann (2013), Li, Q *et al.* (2014), Gunduz e Cataltepe (2015), Khadjeh Nassirtoussi *et al.* (2015), Ingle e Deshmukh (2016), Dang e Duong (2016) e Duong, Nguyen e Dang (2016) foi utilizado o esquema TF-IDF, o valor *term frequency–inverse document frequency* (TF-IDF), que significa frequência do termo–inverso da frequência nos documentos. Trata-se de uma medida estatística que tem a finalidade de indicar a relevância de uma palavra de um documento em relação a

um conjunto de documentos. Já Schumaker *et al.* (2012), Geva e Zahavi (2014), Li, Q. *et al*. (2014) utilizaram o modelo binário em que o peso da característica é 1 se a palavra correspondente estiver presente no documento ou zero se não estiver. Junqué de Fortuny *et al.,* (2014), Li, Q. *et al.* (2014), Alostad e Davulcu (2015), Pröllochs, Feuerriegel, Neumann (2015), Mukund *et al.* (2016), Attigeri (2016), Wei *et al*. (2017) também utilizaram o modelo binário, porém os valores são referentes à polaridade do sentimento extraída do documento, sendo zero para o sentimento negativo e 1 para o sentimento positivo.

Questão 02 - Quais são os classificadores mais utilizados para a identificação de tendências no mercado de ações utilizando-se notícias publicadas na internet?

Por meio do desenvolvimento da revisão sistemática, verificou-se que 69% dos estudos (20/29) utilizaram algoritmos de aprendizagem de máquina. Nos estudos analisados verifica-se que os algoritmos mais utilizados são o *Suport Vector Machine* (SVM), que foi utilizado em 38% dos trabalhos, e o *Naive Bayes, empregado* em 24% dos trabalhos, como mostra a FIG. 12. Deve-se levar em conta que alguns autores utilizaram mais de um algoritmo, como é o caso de Nizer e Nievola (2012), Geva e Zahavi (2014) e Khadjeh Nassirtoussi *et al.* (2015), que usaram tanto o algoritmo SVM como o *Naive Bayes*. Khadjeh Nassirtoussi *et al.* (2015) testaram os algoritmos SVM, *Naive Bayes* e K-NN; Mukund *et al.* (2016) testaram os algoritmos *Naive Bayes*, *Decision Tree* e *Random Tree*. Nos estudos de Li Im *et al*. (2013), Abdullah, Rahaman e Rahman (2013), Kim, Jeong e Ghani (2014) e Wei *et al*. (2017) não foram utilizadas técnicas de aprendizagem de máquina, mas sim as baseadas em léxico, que não dependem de treinamento e não estão sujeitas a serem utilizadas somente em um domínio especifico. Estas dependem de um grupo de termos conhecidos e utilizam métodos estatísticos ou semânticos.

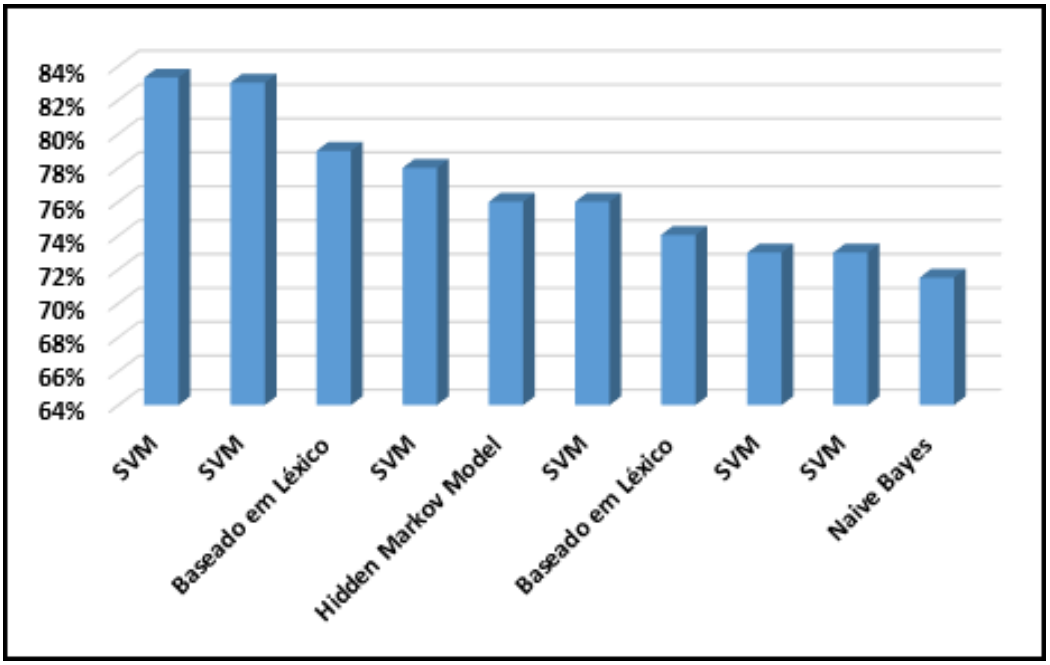
Gráfico 4 - Algoritmos mais utilizados



Fonte: autor.

O algoritmo SVM foi o mais utilizado e também o que retornou melhor precisão entre os trabalhos analisados, como se pode visualizar na FIG. 13. Ele foi aplicado por Nizer e Nievola (2012), Schumaker *et al.* (2012), Fan e Watanabe (2012), Hagenau, Liebmann e Neumann (2013), Geva e Zahavi (2014), Junqué de Fortuny *et al.* (2014), Li, X. (2014), Li, Q. *et al*. (2014), Khadjeh Nassirtoussi *et al.* (2015), Dang e Duong (2016), Duong, Nguyen e Dang (2016), Ichinose e Shimada (2016). O SVM é um classificador linear binário não probabilístico usado para a aprendizagem supervisionada. A principal ideia dos SVMs é encontrar um hiperplano que separe duas classes com uma margem máxima. O problema do treinamento utilizando SVM pode ser representado como um problema de *performance,* dependendo do tamanho da base.

Gráfico 5 - As 10 maiores precisões por algoritmo



Fonte: dados da pesquisa.

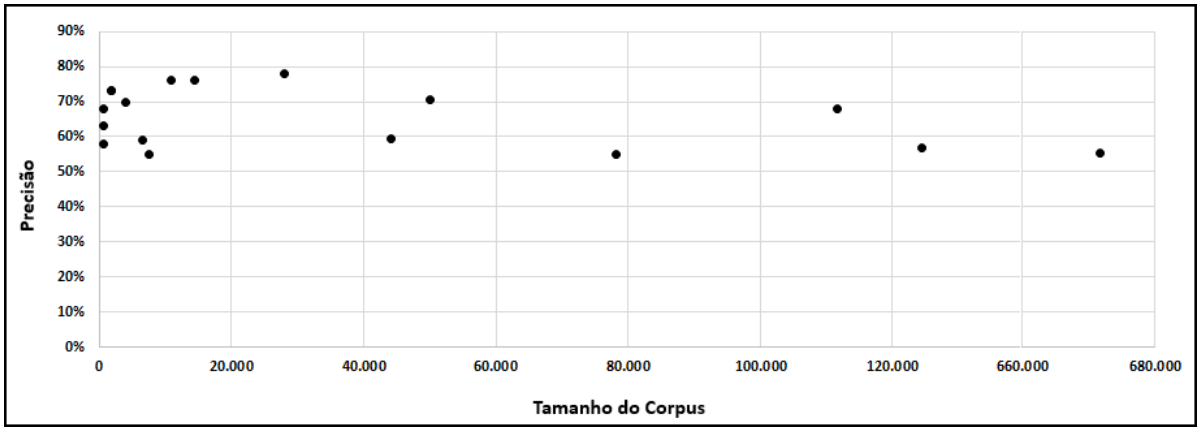
O algoritmo *Naive Bayes* foi o segundo mais utilizado. Ele foi aplicado nos trabalhos de Nizer e Nievola (2012), Geva e Zahavi (2014), Khadjeh Nassirtoussi *et al.* (2014), Gunduz e Cataltepe (2015), Bhardwaj *et al*. (2015), Khadjeh Nassirtoussi *et al.* (2015), Mukund *et al.* (2016). Provavelmente é o algoritmo de classificação mais antigo, mas ainda é muito popular e é usado em muitas das obras. Baseia-se no teorema de Bayes e é chamado de ingênuo porque se calca na suposição ingênua de independência completa entre os recursos de texto. Ele se diferencia de abordagens como *k-Nearest Neighbors* (k-NN), Redes Neurais Artificiais (ANN) ou *Support Vector Machine* (SVM), na medida em que tem como princípio as probabilidades enquanto que nos outros as abordagens mencionadas interpretam a matriz do recurso do documento espacialmente.

Questão 03 - Qual a relação entre o tamanho do corpus e a precisão retornada por um classificador?

Baseado nos estudos analisados, verifica-se que não ocorreu relação direta entre o tamanho do *corpus* e a precisão retornada pelos classificadores utilizados nos experimentos, nos trabalhos de Fan e Watanabe (2012), Hagenau, Liebmann e

Neumann (2013), Alostad e Davulcu (2015), Pröllochs, Feuerriegel, Neumann (2015), Duong, Nguyen e Dang (2016) e Attigeri (2016), que obtiveram precisão igual ou superior a 70%. O tamanho dos *corpora* variou entre 1.884 e 50.000. Já nos trabalhos de Schumaker *et al.* (2012), Kim, Jeong e Ghani (2014)) Junqué de Fortuny *et al.* (2014), Li, Q. *et al*. (2014), Gunduz e Cataltepe (2015), Ichinose e Shimada (2016) e Tirea e Negru (2016), que encontraram precisão abaixo de 70%, o tamanho dos *corpora* variou entre 600 e 671.000. No GRÁF. 6 tem-se a relação entre a precisão e o tamanho dos *corpora* utilizados nos estudos analisados.

Gráfico 6 - Precisão por tamanho do *corpus*



Fonte: dados da pesquisa.

4.4 Conclusão da revisão sistemática da literatura

Revisou-se a literatura focando as técnicas utilizadas na identificação de tendências no mercado de ações por meio de notícias publicadas na internet. Verificou-se que todos os trabalhos utilizaram pelo menos duas fontes de dados como entrada: os dados de texto de notícias extraídas da internet e os dados das cotações das ações. A maioria dos trabalhos usou notícias financeiras por ter menos ruído em comparação às notícias gerais. Dessas notícias foram extraídos os textos principais ou os títulos das notícias os quais também se pode chamar de manchetes. As manchetes das notícias foram utilizadas por serem diretas e, portanto, conter menos ruído comparado com os textos completos e detalhados (GEVA; ZAHAVI, 2014; KHADJEH NASSIRTOUSSI *et al.,* 2015; MUKUND *et al.*, 2016; TIREA; NEGRU, 2016).

A outra fonte de dados de entrada foram as cotações dos preços das ações. Constatou-se que as técnicas de pré-processamento mais citadas foram a remoção de *stop words*, redução das palavras utilizando técnicas de *stemming*, remoção de ruídos que geralmente incluem *tags html*, *links* e propagandas e a tokenização e que uma etapa importante do pré-processamento é a seleção de características do texto, pois o número de palavras diferentes de um texto pode ser grande, mesmo em documentos relativamente pequenos, como *feeds* de notícias. As características mais utilizadas foram a representação por TF-IDF, representação por polaridade do sentimento e a representação binária. Em relação aos classificadores, o algoritmo SVM foi o mais utilizado e também o que retornou melhor precisão entre os trabalhos analisados, seguido do *Naive Bayes*.

Com isso, concluiu-se que os padrões de pré-processamento são similares e o que difere entre os trabalhos analisados foram os classificadores de aprendizado de máquina utilizados e as características escolhidas por esses classificadores. As características usadas nos trabalhos foram na sua maioria representação por polaridade do sentimento utilizando somente dois atributos, sendo o sentimento positivo ou negativo e a representação binária, nesse caso se uma palavra está presente ou não. Acredito que esse modelo pode ser melhorado adicionando-se novos atributos como características, tais como sentimentos que vão além do positivo e negativo.

5 METODOLOGIA

Neste capítulo são apresentados a caracterização da pesquisa e o procedimento metodológico utilizado para responder ao problema de pesquisa e atingir os objetivos levantados para o presente trabalho.

5.1 Caracterização da pesquisa

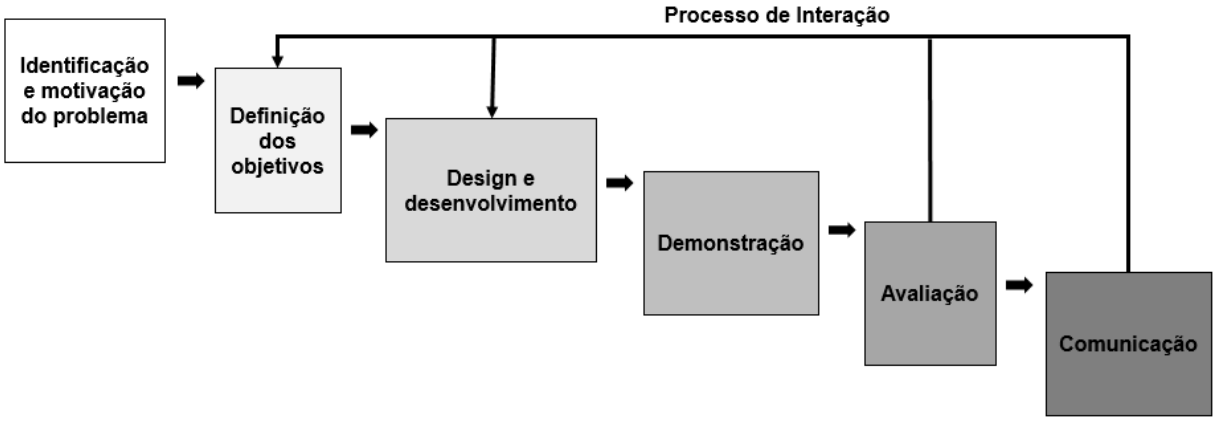
Esta pesquisa caracteriza-se como de natureza aplicada, pois tem o objetivo de estudar técnicas e métodos para identificação de tendências no mercado de ações por meio de notícias publicadas na internet. A abordagem é quantitativa, pois utilizou métodos estatísticos e matemáticos para avaliar a acurácia e precisão dos classificadores de aprendizado de máquina que serão testados no modelo proposto. A técnica de pesquisa é experimental, uma vez que testou diferentes classificadores de aprendizado de máquina. E o método de pesquisa seguiu o método *Design Science Research* (DSR) que, segundo Hevner *et al.,* (2004), baseia e operacionaliza a direção da pesquisa quando o objetivo a ser alcançado é um artefato ou uma prescrição. Como método de pesquisa orientado à solução de problemas, o DSR procura a partir da compreensão do problema construir e avaliar artefatos que possam transformar situações, transformando suas condições para posições melhores ou desejáveis. Ela é empregada na pesquisa como forma de atenuar o distanciamento entre teoria e a prática. Quanto ao rigor e relevância da pesquisa, justificam-se os aspectos específicos deste trabalho mediante a *design science research methodology* (DSRM) proposta por Peffers *et al.* (2007).

5.2 A metodologia proposta por Peffers aplicada à pesquisa

Segundo Peffers *et al.* (2007), o objetivo foi projetar uma metodologia que servisse como um quadro comumente aceito para a realização de pesquisas com base nos princípios de pesquisa do *Design Science* que, em vez de se concentrar nas diferenças de pontos de vista sobre o *Design Science,* entre vários pesquisadores, buscou-se usar uma abordagem de construção de consenso no qual

foi importante para garantir a base da DSRM em elementos bem-aceitos. A FIG. 9 apresenta os elementos do processo.

Figura 9 - *Design Science Research Methodology*



Fonte: adaptada de Peffers *et al.* (2007).

O processo é composto por seis atividades em uma sequência nominal conforme apresentado no QUADRO 2.

Quadro 2 - Atividades do processo proposto por Peffers

Atividades do processo proposto por Peffers *et al.* (2007)	**Aplicação na pesquisa**
1 - Identificação e motivação do problema	Definir o problema de pesquisa específico e justificar o valor de uma solução. A identificação e motivação do problema foram definidos na seção 1.2 que relata a lacuna a ser explorada e na seção 1.4 que relata o problema de pesquisa.
2 - Definição dos objetivos	Inferir os objetivos de uma solução a partir da definição do problema e conhecimento do que é possível e viável. Os objetivos foram definidos na seção 1.5.1 com o objetivo geral e na seção 1.5.2 com os objetivos específicos.
3 - *Design* e desenvolvimento	Criar os artefatos da solução.
4 - Demonstração	Demonstrar o uso do artefato para resolver uma ou mais instâncias do problema.
5 - Avaliação	Observar e medir quão bem o artefato é compatível com uma solução para o problema.
6 - Comunicação	Comunicar a eficácia para os pesquisadores e outros públicos relevantes, como praticantes profissionais, quando apropriado.

Fonte: autor.

5.3 *Design* e desenvolvimento

Nessa atividade criaram-se os artefatos durante o projeto e a implementação da pesquisa. Para Hevner *et al.* (2004) tais artefatos são potencialmente construções, modelos, métodos ou novos recursos técnicos, sociais ou informativos. Conceitualmente, um artefato de pesquisa pode ser qualquer objeto projetado que contribua com o projeto. Nesse projeto considera-se como artefatos o modelo de entidade e relacionamento que representa a base de dados para armazenamento dos dados que foram utilizados durante o desenvolvimento da pesquisa, o *corpus* de notícias gerado sobre a Petrobras, os algoritmos que foram desenvolvidos durante a realização da pesquisa e os arquivos de treinamento.

No desenvolvimento desta pesquisa realizou-se um experimento utilizando-se o *corpus* de notícias criado sobre a Petrobras. No experimento para extrair os sentimentos do texto foi utilizada a API do IBM *Watson Tone Analyzer,* que retorna os índices dos sentimentos alegria, medo, tristeza, raiva e desgosto. Na seção 2.5.3 é descrito em detalhes o funcionamento dessa API. Ela foi selecionada devido à quantidade de sentimentos que ela detecta em um texto, que são cinco. Esses sentimentos são utilizados como características em nosso arquivo de treinamento, diferentemente das outras APIs analisadas, que retornam somente uma característica onde quanto menor o valor, mais negativo o sentimento; e quanto maior o valor, mais positivo o sentimento.

Com isso, criou-se um arquivo de treinamento para o experimento contendo os índices de emoções retornados pela API da IBM como características. O arquivo de treinamento possui como classe os valores -1 e 1, sendo que a classe -1 representa notícias que tiveram seus sentimentos extraídos pela APIs na data em que o fechamento da ação PETR4 foi negativo; e a classe 1 representa notícias que tiveram seus sentimentos extraídos pela APIs na data em que o fechamento da ação PETR4 foi positivo.

5.3.1 Ferramentas utilizadas no desenvolvimento do trabalho

Este tópico faz um breve relato sobre as principais ferramentas e bibliotecas que foram utilizadas no desenvolvimento deste trabalho. A maioria dos itens

descritos a seguir constituem ferramentas gratuitas. As que não são 100% gratuitas possuem algum tipo de limitação que, porém, não impactou o desenvolvimento da pesquisa.

a) *Visual Studio Community*

O *Visual Studio Community* é um ambiente de desenvolvimento gratuito para a criação de aplicativos para *Android*, iOS e *Windows*, bem como aplicativos *web* e serviços na nuvem (MICROSOFT1, 2017). Ele foi utilizado principalmente para o desenvolvimento do aplicativo extrator de notícias e o aplicativo para a criação dos arquivos de treinamento.

b) *SQL Server Express Edition*

O *SQL Server Express Edition* é um banco de dados de nível básico gratuito onde se pode criar base de dados de até 10 GB (MICROSOFT2, 2017). Utilizou-se esse banco de dados para armazenar o *corpus* de notícias extraídas da internet e demais dados que foram armazenados nas tabelas de apoio que estão descritas no QUADRO 4.

c) WEKA

WEKA é uma aplicativo que contém uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados e textos. Os algoritmos podem ser aplicados diretamente a um conjunto de dados ou chamados a partir de seu próprio código Java. WEKA é uma ferramenta de código aberto emitido sob a *General Public License* (GNU) e foi desenvolvido pela Universidade de Waikato da Nova Zelândia.

A ferramenta WEKA foi utilizada como *framework* dos classificadores utilizados nos experimentos como também a ferramenta para avaliação dos resultados retornados pelo arquivo de treinamento (WAIKATO, 2017).

d) *API do IBM Watson Tone Analyzer*

Esse serviço utiliza análise linguística para detectar e interpretar emoções, tendências sociais e pistas do estilo de linguagem encontradas no texto. As

emoções detectadas incluem alegria, medo, tristeza, raiva e desgosto (IBM1, 2018a). A utilização da API é gratuita para até 50 mil requisições por mês.

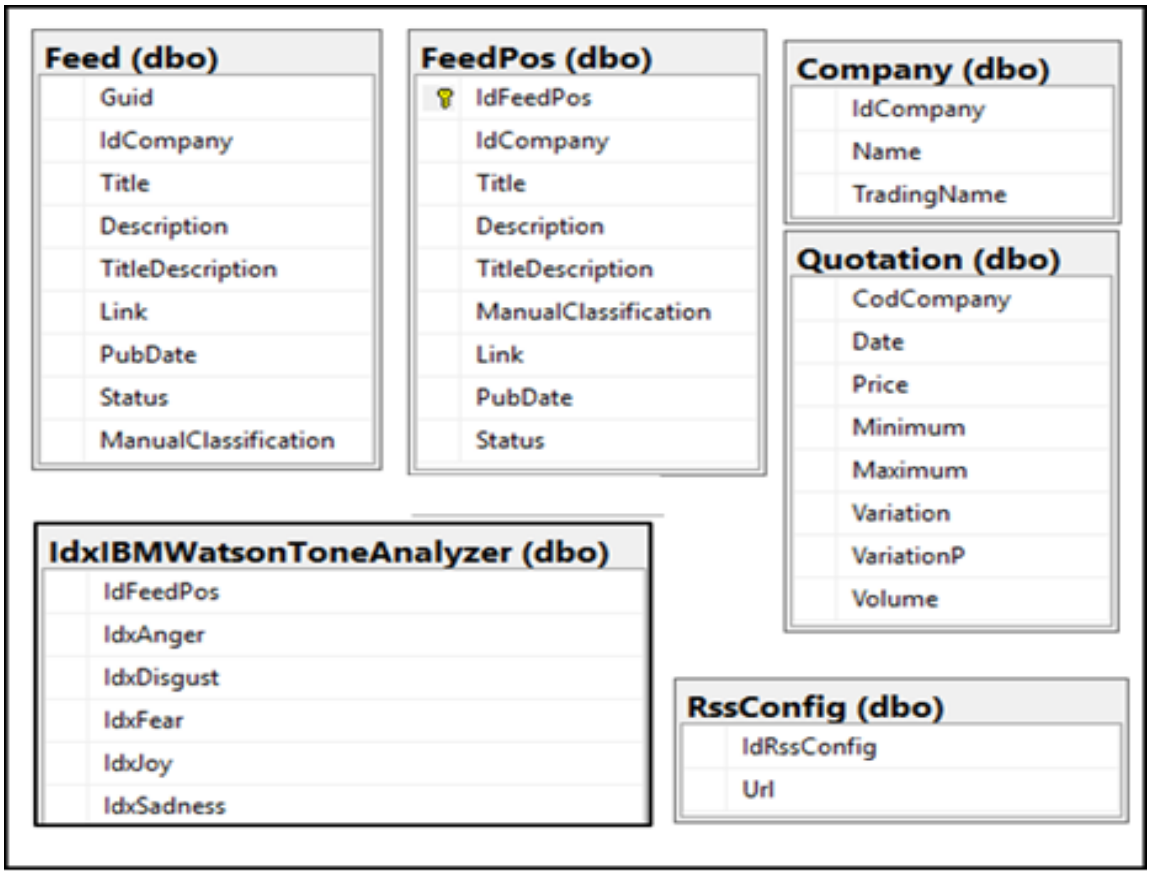
e) *API do IBM Watson Language Translator*

Permite a tradução de textos em domínios de notícias, conversas e patentes. Utilizou-se essa API para traduzir os *feeds* de notícias antes de submetê-las à extração de emoções pelo *IBM Watson Tone Analyzer* (IBM2, 2018b). A utilização da API é gratuita para até 50 mil requisições por mês.

5.3.2 Criação da base de dados

A primeira atividade de *design* e desenvolvimento realizada foi a criação da base de dados com todas as tabelas de apoio para armazenamento dos dados necessários para o desenvolvimento da pesquisa. Na FIG. 10 tem-se o diagrama de entidades de todas as tabelas utilizadas na pesquisa.

Figura 4 - Modelo de entidades



Fonte: autor.

No QUADRO 3 tem-se a descrição de todas as tabelas que foram utilizadas na implementação do projeto de pesquisa.

Quadro 3 - Descrição das tabelas do modelo de entidade e relacionamento

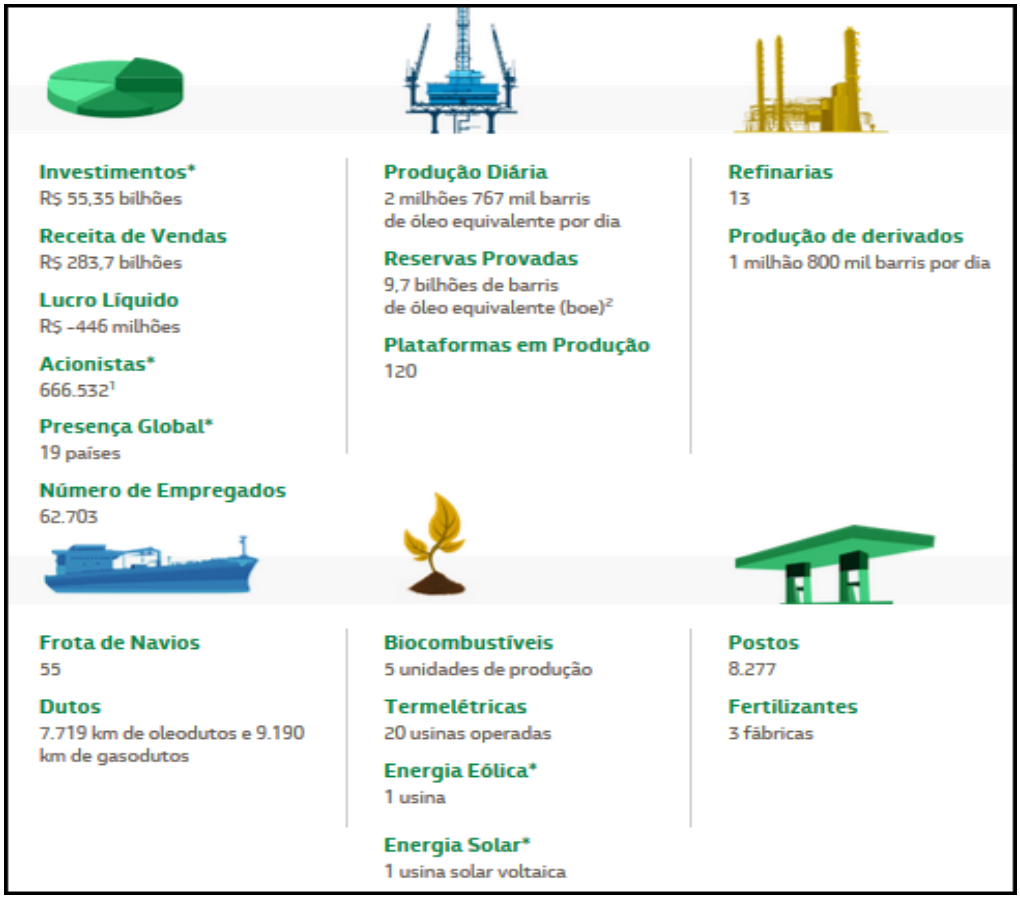
Tabela de Apoio	Descrição
RssConfig	Tabela que armazena os endereços RSS dos *sites* de notícias utilizados pelo processo de extração dos *feeds*.
Quotation	Tabela de apoio onde são armazenadas as cotações diárias do papel PETR4.
Company	Tabela onde são armazenado o nome das empresas das quais quer-se extrair as notícias, utilizamos no trabalho somente a empresa Petrobras.
Feed	Tabela onde são armazenados os *feeds* de notícias das empresas cadastradas na tabela *Company*. Nesta tabela as notícias não sofrem tipo algum de processamento, sendo armazenado o conteúdo no seu formato original.
FeedPos	Tabela onde são armazenados os *feeds* de notícias depois de passar pela etapa de pré-processamento.
IdxIBMWatsonToneAnalyzer	Tabela onde serão armazenados os índices de emoções extraídas das notícias por meio da *API do IBM Watson Tone Analyzer*.

Fonte: autor.

5.3.3 Seleção da empresa para realização dos experimentos

A empresa utilizada nesta pesquisa foi a Petróleo do Brasil S.A., que possui o nome de pregão "Petrobras" na BM&F Bovespa e o papel (ação) utilizado foi o PETR4. A escolha da empresa Petrobras para realização da pesquisa se deu por sua importância no mercado nacional e o volume do movimento diário de compra e venda de suas ações. Segundo estudos da Economatica (2018), a ação da Petrobras PN (PETR4) é a que tem maior volume financeiro médio diário nos últimos 10 anos, considerando o período de 15 de fevereiro de 2008 até 15 de fevereiro de 2018, com média de R$ 615,0 milhões/dia. A Petrobras foi fundada em 3 de outubro de 1953, é uma empresa de sociedade anônima de capital aberto que atua de forma integrada e especializada na indústria de óleo, gás natural e energia. Está presente nos segmentos de exploração, produção, refino, comercialização, transporte, petroquímica, distribuição de derivados, gás natural, energia elétrica, gás-química e biocombustíveis. O perfil da empresa é apresentado na FIGURA 17.

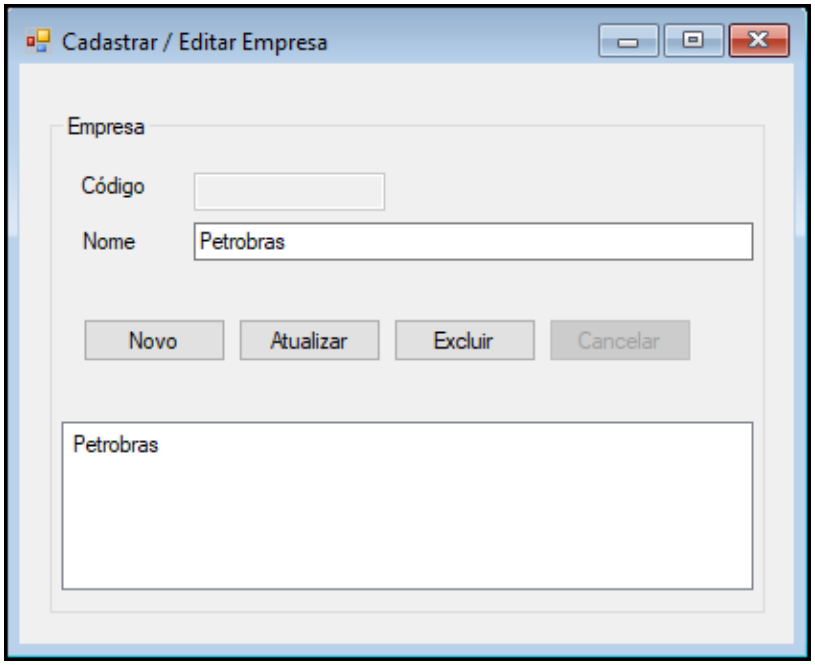
Figura 11 - Perfil da empresa Petrobras



Autor: Petrobras3 (2017).

Na implementação da pesquisa cadastrou-se na tabela “*Company*” o nome Petrobras por meio de uma aplicação *Windows Form* desenvolvida na linguagem de programação C#. De acordo com a FIG. 12, esse campo é utilizado no processo de extração dos *feeds* de notícias para identificar notícias que citam a Petrobras.

Figura 12 - Interface de cadastro e edição das empresas



Fonte: autor.

5.3.4 Seleção dos sites de notícias

Na etapa de seleção dos *sites* de notícias foram escolhidos *sites* de *feeds* de notícias que disponibilizam seu conteúdo de notícias por *Really Simple Syndication* (RSS).

RSS é frequentemente chamado de *feed* de notícias ou *feed* RSS. Trata-se de uma tecnologia que facilita a distribuição de conteúdo, definindo uma maneira fácil de compartilhar e visualizar títulos e conteúdo. Os arquivos RSS podem ser atualizados automaticamente e permitem visualizações personalizadas para diferentes *sites*. Com o RSS é possível distribuir conteúdo da *web* atualizado de um *site* para milhares de outros *sites* ou programas que reúnem e classificam *feeds* RSS, também conhecidos como agregadores RSS (HARVARD LAW, 2003; W3SCHOOLS, 2018).

RSS é útil para *sites* que são atualizados com frequência, como *sites* de notícias, permitindo aos usuários a navegação rápida por notícias e atualizações. RSS foi projetado para mostrar dados selecionados. Sem ele, os usuários terão que verificar seu *site* diariamente para novas atualizações, podendo consumir muito tempo dos usuários. Com um *feed* RSS, usuários podem verificar seu *site* mais rapidamente usando um agregador de *feeds* RSS. Os documentos RSS usam

sintaxe simples e autodescritiva (SINGH; SAHU, 2015). Segue na FIG. 13 um exemplo de um documento RSS simples.

Figura 13 - Exemplo de um documento RSS simples

```
<?xml version="1.0" encoding="UTF-8" ?>
<rss version="2.0">

<channel>
  <title>W3Schools Home Page</title>
  <link>https://www.w3schools.com</link>
  <description>Free web building tutorials</description>
  <item>
    <title>RSS Tutorial</title>
    <link>https://www.w3schools.com/xml/xml_rss.asp</link>
    <description>New RSS tutorial on W3Schools</description>
  </item>
  <item>
    <title>XML Tutorial</title>
    <link>https://www.w3schools.com/xml</link>
    <description>New XML tutorial on W3Schools</description>
  </item>
</channel>

</rss>
```

Fonte: W3SCHOOLS (2018).

A primeira linha do documento define a declaração XML, define a versão XML e a codificação de caracteres usada no documento. Nesse caso, o documento está em conformidade com a especificação 1.0 do XML e usa o conjunto de caracteres UTF-8. A próxima linha é a declaração RSS, que identifica que esse é um documento RSS, nesse caso, RSS versão 2.0.

A próxima linha contém o elemento <channel>. Esse elemento é usado para descrever o *feed* RSS.

O elemento <channel> tem três elementos filhos obrigatórios:

a) <title> - define o título do canal
b) <link> - define o hiperlink para o canal

c) <description> - descreve o canal

Cada elemento <channel> pode ter um ou mais elementos <item>.
Cada elemento <item> define um artigo no *feed* RSS.
O elemento <item> possui três elementos filhos obrigatórios:
<title> - define o título do item
<link> - define o hiperlink para o item
<description> - o conteúdo do item
Finalmente, as duas últimas linhas fecham os elementos <channel> e <rss>.

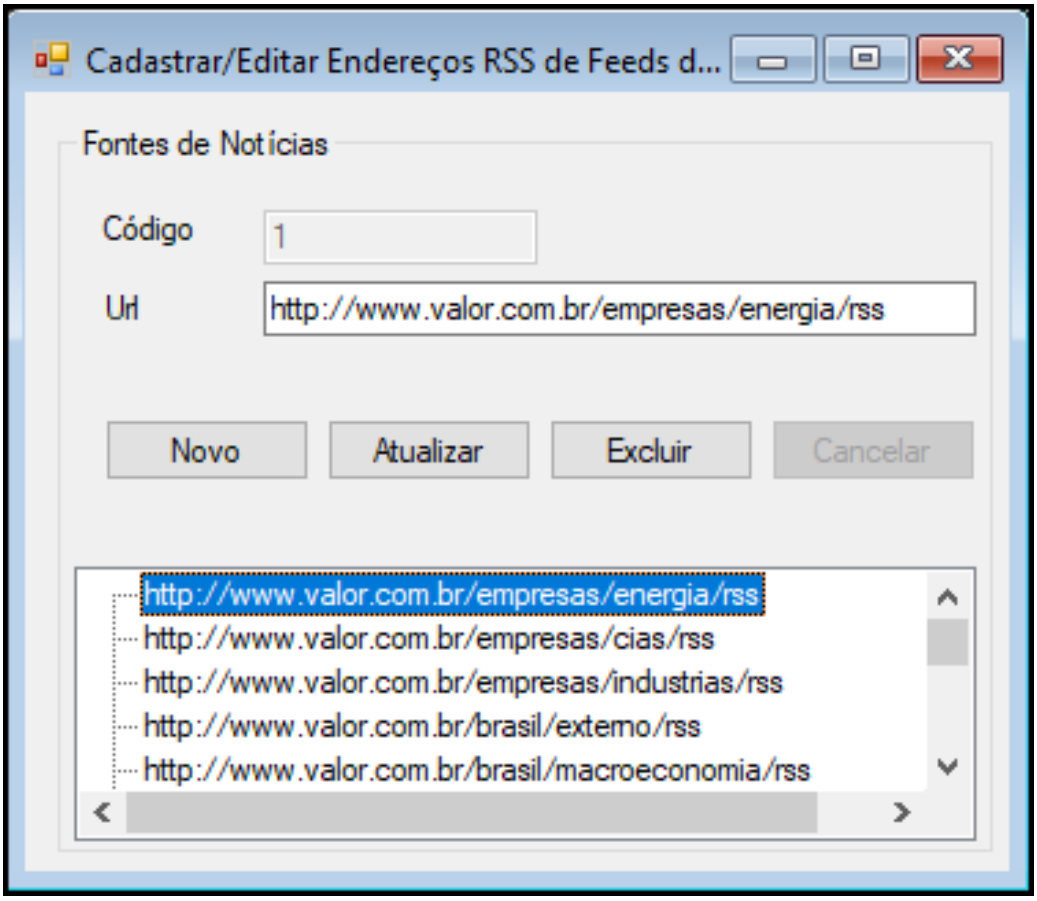
Os arquivos RSS são acessados por meio de uma *Uniform Resource Locator* (URL) e cada item do arquivo corresponde a um *feed* de notícia que possui um título, que pode também ser chamado de manchete, uma descrição que corresponde ao conteúdo da notícia, data de publicação e o *link* original da notícia.

Utilizaram-se para a pesquisa as seguintes *URLs* de *feeds* RSS:

a) http://www.valor.com.br/empresas/energia/rss
b) http://www.valor.com.br/empresas/cias/rss
c) http://www.valor.com.br/empresas/industrias/rss
d) http://www.valor.com.br/brasil/externo/rss
e) http://www.valor.com.br/brasil/macroeconomia/rss
f) http://g1.globo.com/dynamo/brasil/rss2.xml
g) http://g1.globo.com/dynamo/economia/rss2.xml
h) http://www.em.com.br/rss/noticia/economia/rss.xml
i) http://www.infomoney.com.br/negocios/noticias-corporativas/rss
j) http://www.infomoney.com.br/bloomberg/rss
k) http://www.infomoney.com.br/petrobras/rss

Cada endereço URL dos *feeds* de notícias utilizados na pesquisa foram cadastrados na tabela “RssConfig” por meio de uma aplicação *Windows Form* desenvolvida na linguagem de programação C#. Na FIG. 14 tem-se a imagem da interface da aplicação para o cadastro e edição das URLs para os endereços RSS dos *feeds* de notícias.

Figura 14 - Interface de cadastro e edição das URLs dos endereços RSS dos *feeds* de notícias



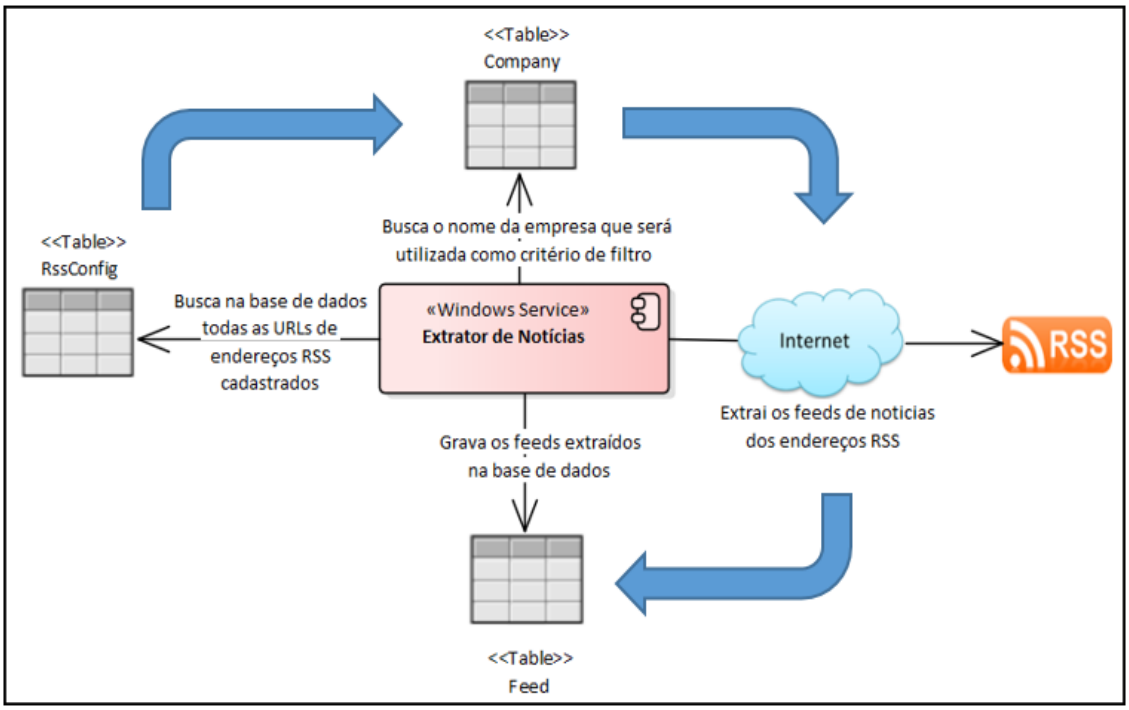
Fonte: autor.

As URLs cadastradas na tabela “RssConfig” serão utilizadas no processo de extração e criação do *corpus* de notícias.

5.3.5 Extração e criação do corpus de notícias e cadastramento do fechamento diário do papel PETR4

Com os *sites* de notícias e seus endereços RSS cadastrados na base de dados na tabela “RssConfig” a extração das notícias foi feita por um aplicativo *Windows Service* construído na linguagem de programação C#.NET e executado como um processo do *Windows* que tem como característica rodar em segundo plano no sistema operacional, não possuindo uma interface gráfica. E de 10 em 10 segundos consulta cada endereço URL dos arquivos RSS cadastrados na base de dados na tabela “RssConfig” para verificar se existem novas notícias. Caso positivo, o serviço extrai os *feeds* de notícias e verifica se no título ou conteúdo da notícia possui a palavra “Petrobras” que foi cadastrada na tabela “*Company*”. Se positivo, esse *feed* é armazenado no banco de dados na tabela “Feed”. As notícias não se repetem, pois cada *feed* possui uma chave única que é validada antes de salvá-lo no banco de dados. Segue na FIG. 15 a representação do componente que realiza a extração e armazenamento dos *feeds* de notícias.

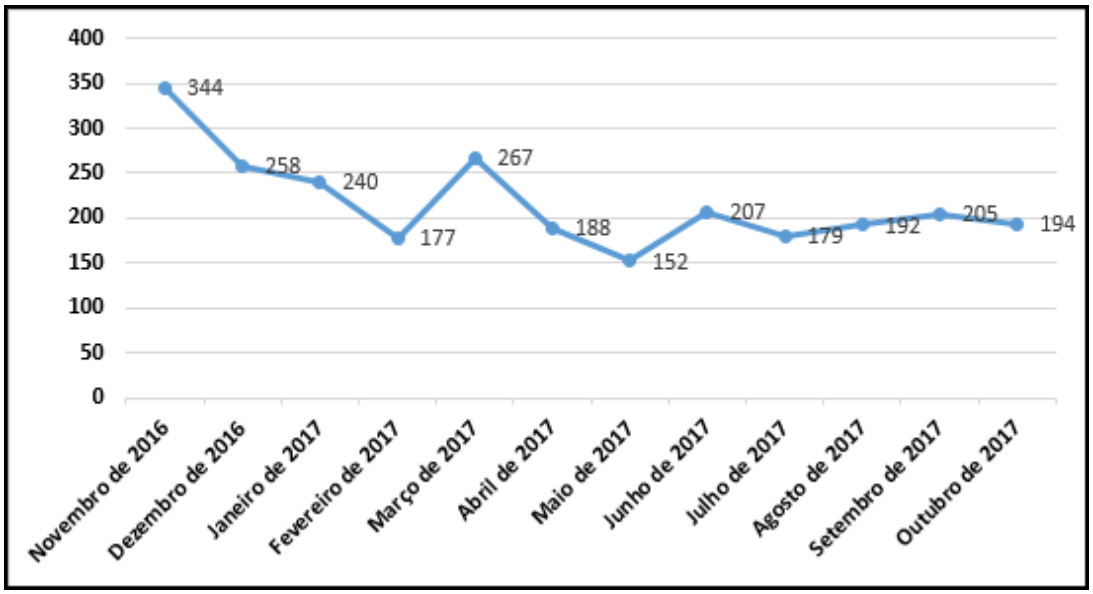
Figura 15 - Componente de extração e armazenamento dos *feeds* de notícias



Fonte: autor.

Foram extraídos 2.603 *feeds* no período entre 1º/11/2016 e 31/10/2017 referentes à Petrobras. Cada *feed* possui, em média, 242 caracteres considerando-se a manchete e o conteúdo do *feed,* que equivale a 242 *bytes* por *feed* e 629,926 *Kbytes* no total do *corpus*. Segue no GRÁF. 7 a quantidade de *feeds* de notícias extraídas por mês.

Gráfico 7 - Quantidade de notícias extraídas por mês



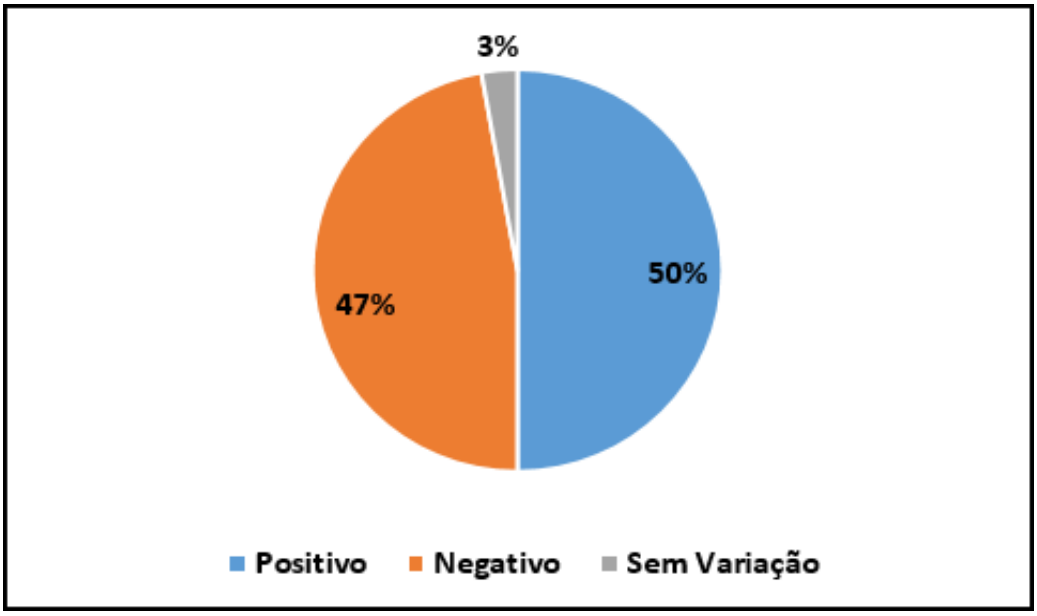
Fonte: autor.

O fechamento das cotações diárias do papel PETR4 da empresa Petrobras foi extraído manualmente da página do UOL economia, que pode ser acessada pelo endereço https://economia.uol.com.br/cotacoes/bolsas/acoes/bvsp-bovespa/petr4-sa/. As cotações foram incluídas na tabela "*Quotation*" descrita no QUADRO 4. Os dados salvos nessa tabela foram: o código da empresa no mercado de ações, no caso desta pesquisa o código é PETR4, referente à empresa Petrobras; a data; o preço com que o papel fechou; o valor mínimo do papel durante o dia; o preço máximo do papel durante o dia; a variação do preço do papel comparado com o fechamento anterior; a variação do papel em percentual comparado com o fechamento anterior; e o volume movimentado no dia.

Foram extraídas as cotações no período de 1º/11/2016 a 31/10/2017. No total foram extraídas 248 cotações referentes ao fechamento diário do papel PETR4, sendo que, do total, 124 cotações fecharam no positivo, 117 cotações fecharam no negativo e sete fecharam sem variação no período. O GRÁF. 8 mostra o

comportamento do fechamento do papel PETR4 no período entre 1/11/2016 e 31/10/2017.

Gráfico 8 - Fechamento diário do papel PETR4 no período de 1/11/2016 a 31/10/2017



Fonte: autor.

5.3.6 Pré-processamento do corpus de notícias

a) Remoção dos ruídos

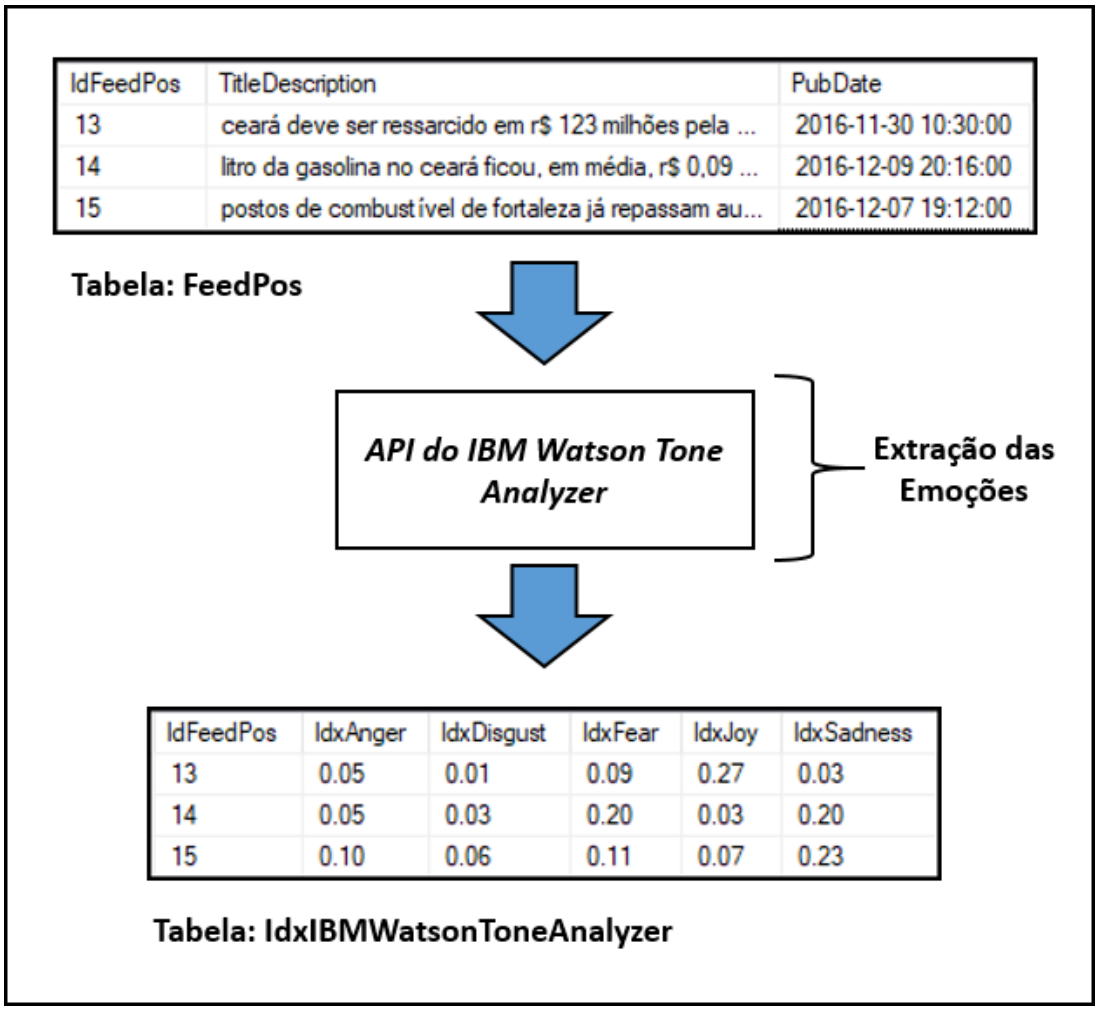
Foram removidos ruídos dos *feeds* de notícias tais como caracteres especiais, TAGs html e propagandas que estavam armazenados na tabela “Feed”. Após a remoção dos ruídos esses *feeds* foram movidos para a tabela “FeedPos”.

b) Extração dos sentimentos utilizando a API do *IBM Watson Tone Analyzer*

Para extrair os índices de emoções dos *feeds* de notícias foi necessário antes traduzir o texto de português para o inglês, pois o IBM *Watson Tone Analyzer* não tem o suporte do idioma português. Foi desenvolvido um aplicativo na linguagem C# .NET responsável por ler cada *feed* de notícia e traduzi-lo para o inglês por meio da API do IBM *WatsonLanguage Translator.*

Após a tradução as emoções foram extraídas de cada *feed* de notícia utilizando a API do IBM *Watson Tone Analyzer*. Ao finalizar a extração dos índices de sentimentos dos *feeds* de notícias, os valores desses índices foram armazenados na tabela IdxIBMWatsonToneAnalyzer, conforme mostrado na FIG. 16.

Figura 16 - Extração dos sentimentos dos *feeds* de notícias por meio da API do *IBM Watson Tone Analyzer*



Fonte: autor.

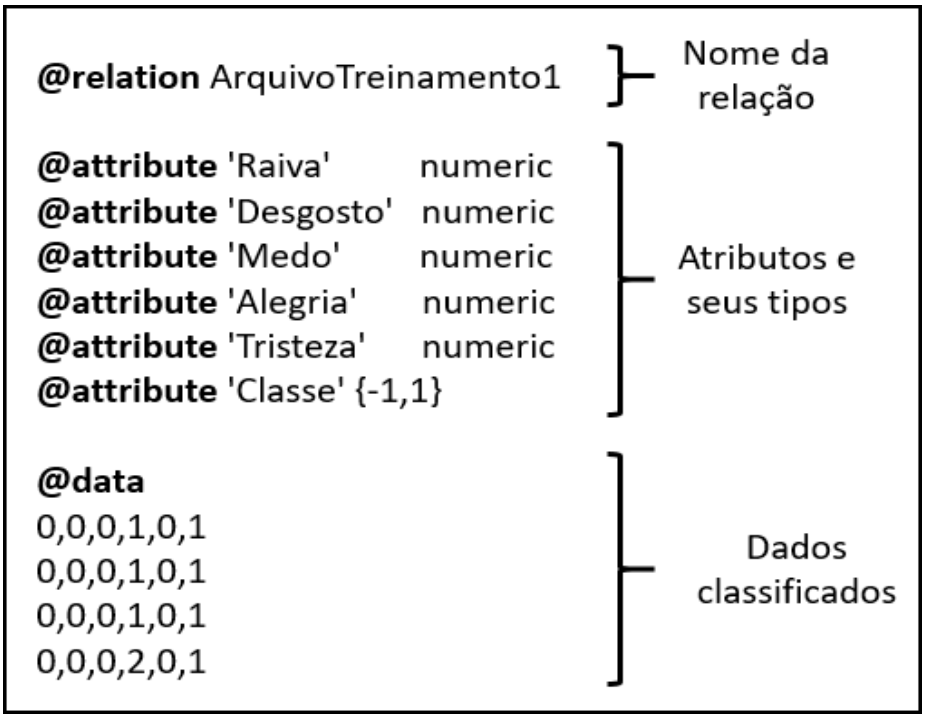
5.3.7 Geração do arquivo de treinamento

O arquivo de treinamento foi criado no formato *Attribute Relation File Format* (ARFF), que é o padrão lido pelo aplicativo Weka. Os arquivos ARFFs possuem duas seções distintas. A primeira são as informações do cabeçalho e a segunda seção são os dados. O cabeçalho do arquivo ARFF contém o nome da relação que

é definida pela *tag* *@RELATION*, uma lista dos atributos definidos pela *tag* *@ATTRIBUTE* (são as colunas dos dados) e seus tipos. Os dados são colocados após a tag *@DATA*.

Um exemplo de arquivo de treinamento no formato ARFF é ilustrado na FIG. 17.

Figura 17 - Exemplo de um arquivo no formato ARFF



Fonte: autor.

O QUADRO 4 exibe a descrição do arquivo de treinamento utilizado neste trabalho.

Quadro 4 - Descrição dos arquivos de treinamento gerados

@Relation	***@Attribute***	***@Data***	**Classe**
ArquivoTreinamento1	Índice_Raiva Índice_Desgosto Índice_Medo Índice_Alegria Índice_Tristeza	Os índices originais foram convertidos para os valores de acordo com as seguintes faixas: Índice < 0,50 = 0 Índice entre 0,50 e 0,75 = 1 Índice > 0,75 = 2	1 para a data que o fechamento da ação for > 0. -1 para a data que o fechamento da ação for < 0.

Fonte: autor.

5.3.7.1 Geração do arquivo de treinamento para o experimento

Para facilitar a criação do arquivo de treinamento do experimento foi criada a *view* “V_IBM_WatsonToneAnalyzer” com o objetivo de unir as informações da tabela “*Quotation”* que contém as cotações do papel PETR4 da empresa Petrobras com a tabela “IdxIBMWatsonToneAnalyzer”, que possui os índices referentes aos sentimentos extraídos pela API do IBM *Watson Tone Analyzer* dos *feeds* de notícias, conforme exibido na FIG. 18.

Figura 18 - *View* resultado da união dos dados da tabela *Quotation* e *IdxIBMWatsonToneAnalyze*r

IdFeedPos	TitleDescription	PubDate	VariationP	IdxAnger	IdxDisgust	IdxFear	IdxJoy	IdxSadness
13	ceará deve ser ...	2016-11-30 10:30:00	9.14	0.05	0.01	0.09	0.62	0.03
14	litro da gasolina ...	2016-12-09 20:16:00	-0.76	0.05	0.03	0.20	0.03	0.20
15	postos de comb...	2016-12-07 19:12:00	-1.80	0.10	0.06	0.11	0.07	0.23
16	camilo busca ira...	2017-01-17 17:50:00	0.44	0.17	0.20	0.15	0.12	0.35
17	postos do ceará...	2017-01-31 15:54:00	1.21	0.04	0.05	0.06	0.06	0.32
18	bovespa sobe, ...	2016-11-04 10:25:00	-0.87	0.15	0.03	0.18	0.04	0.53
19	bovespa: 11.11...	2016-11-11 10:10:00	-9.61	0.05	0.03	0.31	0.02	0.21
20	bovespa cai ma...	2016-11-03 18:18:00	-4.33	0.17	0.02	0.12	0.06	0.57
21	bovespa cai qu...	2016-11-29 18:16:00	-5.17	0.16	0.04	0.23	0.09	0.45
22	bovespa fecha ...	2016-11-21 18:18:00	7.30	0.14	0.05	0.09	0.87	0.22

Fonte: autor.

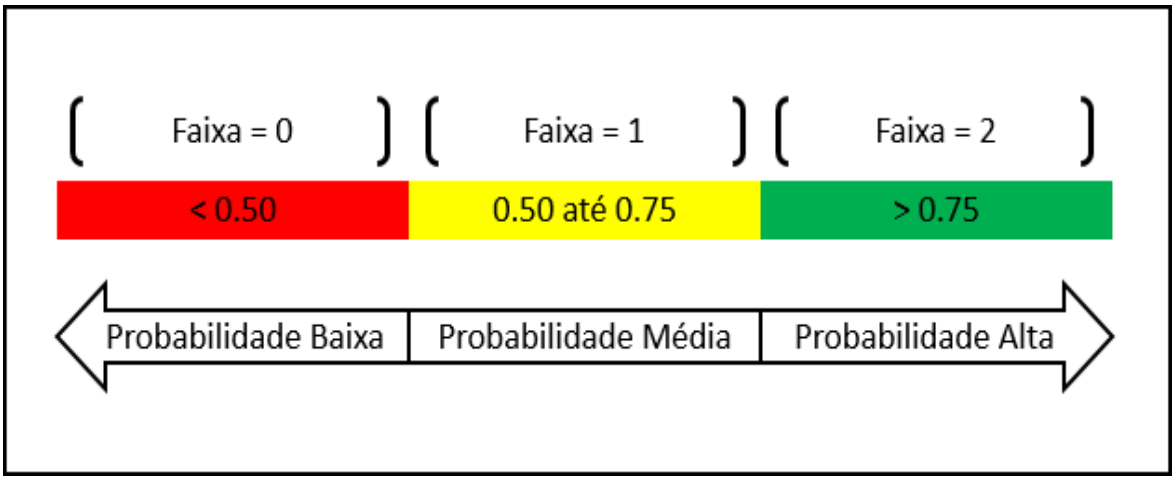
Uma *view* pode ser definida como uma tabela virtual combinada por linhas e colunas de dados originários de tabelas relacionadas em uma *query*. As linhas e colunas de uma *view* são criadas de forma dinâmica no momento que são referenciadas (PICHILIANI, 2018).

Para gerar o arquivo de treinamento foi desenvolvida uma aplicação em C#.NET que lê a *view* V_IBM_WatsonToneAnalyzer, processa os dados e gera o arquivo de ArquivoTreinamento1.arff no formato que pode ser lido pelo aplicativo WEKA. O arquivo de treinamento foi gerado com o seguinte padrão: como nome da relação adotou-se o nome @ArquivoTreinamento1, como atributos usaram-se os campos Indice_Raiva, que representa o campo IdxAnger da view V_IBM_WatsonToneAnalyzer, Indice_Desgosto que representa o IdxDisgust,

Indice_Medo que representa o IdxFear, Indice_Alegria que representa o IdxJoy e o Indice_Tristeza que representa o IdxSadness.

Antes de alimentar os dados do arquivo de treinamento o valor de cada atributo foi convertido para um novo valor de acordo com a faixa em que ele se enquadra. As faixas são definidas pelo algoritmo utilizado para extração dos sentimentos. No caso do experimento utilizou-se a API do IBM *Watson Tone Analyzer,* que retorna três faixas diferentes. A primeira faixa significa que quando um índice retorna um valor menor do que 0,50 o sentimento que esse índice representa tem baixa probabilidade de estar correto. Já a segunda faixa é quando o valor de um índice está entre 0,50 e 0,75 e significa que a probabilidade do sentimento que esse índice representa estar correta é média. Por fim, a terceira faixa, onde os valores dos índices que são maiores que 0,75 têm alta probabilidade do sentimento que o índice representa estar correto. As faixas definidas pela API *do IBM Watson Tone Analyzer* são exibidas na FIG. 19.

Figura 19 - Faixa de probabilidade de um índice estar correto definido pela *API do IBM Watson Tone Analyzer*



Fonte: autor.

Os índices com valores abaixo de 0,50 no arquivo de treinamento foram convertidos para zero; os índices com valores entre 0,50 e 0,75 no arquivo de treinamento foram convertidos para 1; e os índices com valores acima de 0,75 no arquivo de treinamento foram convertidos para dois.

Como classe foi definido que quando o campo “VariationP” da *view* “V_IBM_WatsonToneAnalyzer” fosse menor que zero o valor utilizado na classe

seria o -1; e se for maior que zero, o valor seria 1. O campo VariationP representa o fechamento do papel PETR4 na data dos *feeds* de notícias que foram utilizados, ou seja, sempre que o valor da classe do arquivo de treinamento era igual a -1 significava que o papel PETR4 tinha fechado no negativo; e quando o valor era igual a 1 o papel PETR4 tinha fechado no positivo, sendo que cada linha do arquivo de treinamento representava um *feed* de notícia do qual foram extraídos os índices de sentimentos. Para esta pesquisa desconsideraram-se os *feeds* de notícias na data em que a variação do dia foi igual a zero, pelo fato de a quantidade de registros nessa situação diferir muito da quantidade de registros nas datas com fechamento positivo ou negativo, o que compromete o treinamento com os algoritmos de aprendizado de máquina.

5.4 Demonstração e avaliação do experimento

Nesta atividade é demonstrado o uso dos artefatos a partir de experimentações, no *Design Science Research.* Segundo Dresch, Lacerda e Antunes Júnior (2013), essa avaliação pode ocorrer utilizando simulação computacional, experimentos em laboratórios ou experimentos em campo.

Nesta pesquisa realizaram-se experimentos utilizando o arquivo de treinamento gerado na seção 5.3.7 para avaliar a precisão do uso dos sentimentos extraídos dos *feeds* de notícias, como uma forma para identificar a tendência de alta ou baixa do papel PETR4 no mercado de ações. Para a realização dessa atividade utilizou-se a ferramenta WEKA, que possui uma coleção de classificadores de aprendizado de máquina, dos quais foram utilizados os classificadores KNN, *Naive Bayes* e SVM.

5.4.1 Técnicas de treinamento e validação

a) K-Fold Cross-Validation

Para o treinamento utilizaram-se os classificadores KNN, *Naive Bayes* e SVM. A técnica de treinamento adotada para cada classificador foi a validação cruzada k-*fold*, que é um método estatístico de avaliação e comparação de algoritmos de aprendizagem dividindo dados em dois segmentos: um usado para aprender ou

treinar um modelo e o outro usado para validar o modelo. Na validação cruzada típica os conjuntos de treinamento e validação devem ser cruzados em rodadas sucessivas de modo que cada ponto de dados tenha uma chance de ser validado (RAHMAN *et al.*, 2003; YADAV; SHUKLA, 2016).

b) Precisão e revocação

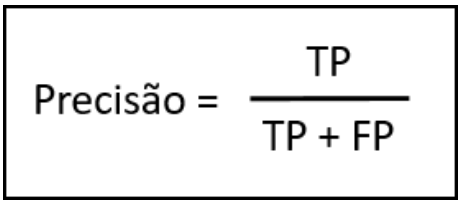
Após a simulação extraíram-se para comparação os resultados de precisão, revocação e a matriz confusão para cada classificador utilizado no experimento. Precisão trata-se da fração de instâncias relevantes entre as instâncias recuperadas (FIG. 20).

Figura 20 - Precisão

Fonte: autor.

Por exemplo, para uma pesquisa de texto em um conjunto de documentos, a precisão é o número de resultados corretos dividido pelo número de todos os resultados retornados (FLACH; KULL, 2015). A fórmula para se calcular a precisão encontra-se na FIG. 21.

Figura 21 - Fórmula para calcular a precisão



Fonte: autor.

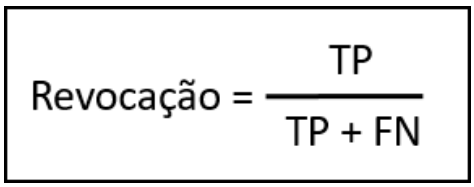
TP indica o número de verdadeiro-positivos e FP indica o número de falso-positivos. Já a revocação é a fração de instâncias relevantes que foram recuperadas ao longo do total de instâncias relevantes, de acordo com a a FIG. 22.

Figura 22 - Revocação

Fonte: autor.

Cita-se como exemplo que, para uma pesquisa de texto em um conjunto de documentos, a revocação é o número de resultados corretos dividido pelo número de resultados que deveriam ter sido retornados. Pode ser vista como a probabilidade de um documento relevante ser recuperado pela consulta (ABDULHAMID; OSHO; ISMAILA, 2018). Segue a fórmula para se calcular a precisão na FIG. 23.

Figura 23 - Fórmula para calcular a revocação



Fonte: autor,

TP indica o número de verdadeiro-positivos e FN o número de falso-negativos.

c) Matriz de confusão

A matriz de confusão é um conceito de aprendizado de máquina que contém informações sobre classificações reais e previstas realizadas por um sistema de classificação. Uma matriz de confusão tem duas dimensões: uma é indexada pela classe real de um objeto e a outra é indexada pela classe que o classificador prevê.

A FIG. 24 apresenta a forma básica da matriz confusão para uma tarefa de classificação multiclasse, com as classes A_1, A_2 e A_n. Na matriz confusão, N_{ij} representa o número de amostras realmente pertencentes à classe Ai, mas classificadas como classe A_j (DENG *et al.,* 2016).

Figura 24 - Matriz confusão

		Previsto		
		A_1	A_j	A_n
Real	A_1	N_{11}	N_{1j}	N_{1n}
	A_j	N_{i1}	N_{ij}	N_{in}
	A_n	N_{n1}	N_{nj}	N_{nn}

Autor: Deng *et al.* (2016).

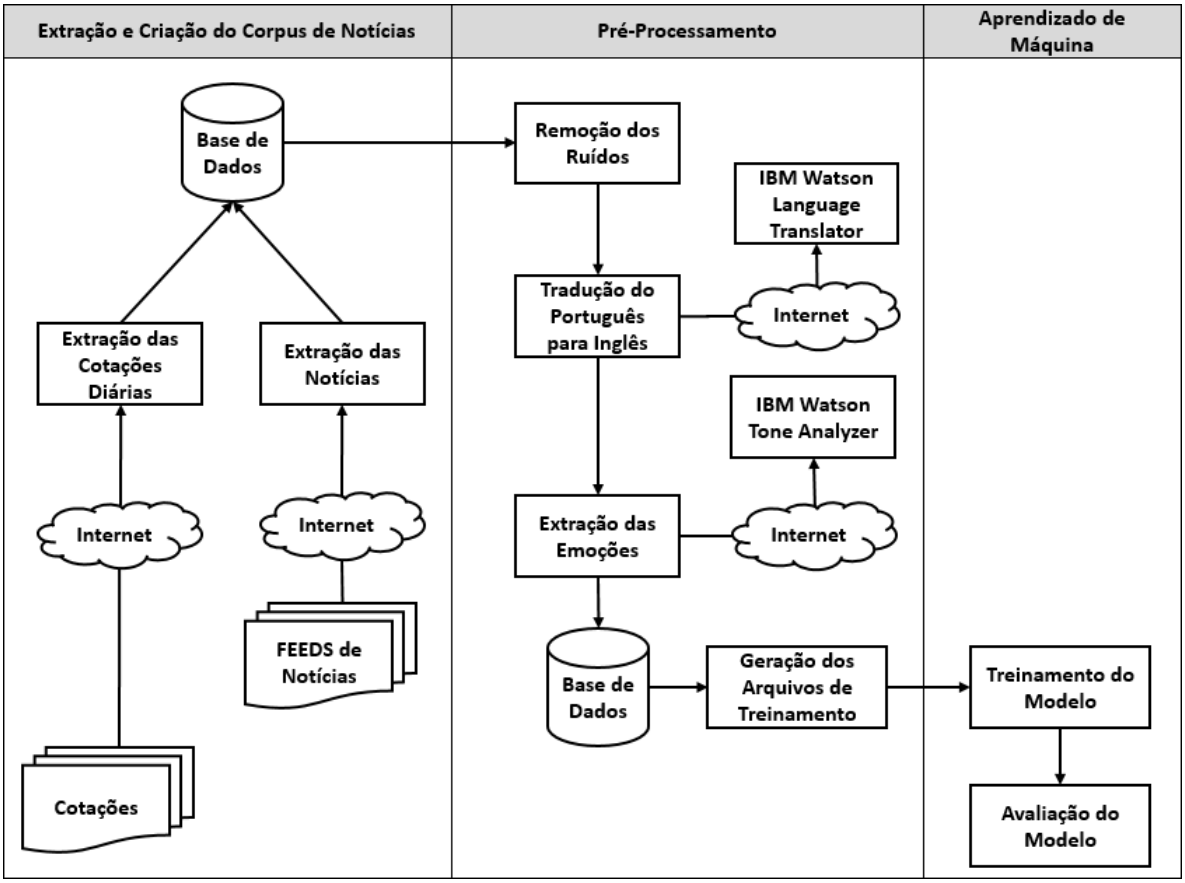
5.4.1 Experimento

No experimento utilizou-se um *corpus* com 2.525 instâncias, em que cada instância representou um *feed* de notícia, sendo que para compor os atributos do arquivo de treinamento foram utilizados os índices de sentimentos extraídos de cada *feed* de notícia pela API *do IBM Watson Tone Analyzer* (raiva, desgosto, medo, alegria, tristeza). No arquivo de treinamento o valor de cada índice foi convertido de acordo com as faixas de valores definidas na seção 5.3.7.1.

Das 2.525 instâncias que foram utilizadas no arquivo de treinamento, 1.197 foram definidas com a classe -1, que representa 47% do total de instâncias, e 1.328 com a classe 1, que representa 53% do total de instâncias. A classe -1 representa as notícias em cuja data de publicação o papel PETR4 fechou no negativo; e a classe 1 representa as notícias em cuja data de publicação o papel PETR4 fechou no positivo.

Na FIG. 25 vê-se o modelo em alto nível da arquitetura da aplicação do experimento.

Figura 25 - Modelo arquitetural em alto nível da aplicação do experimento



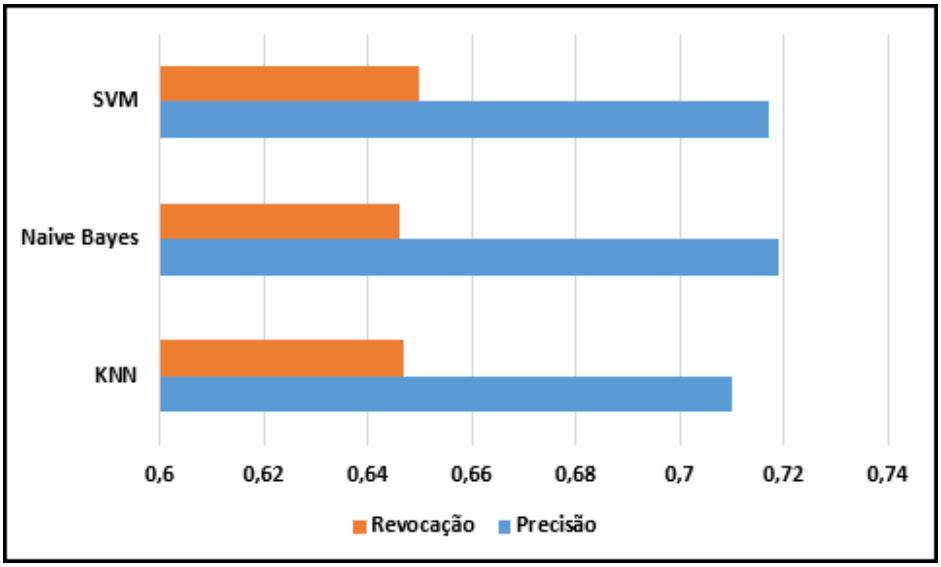
Fonte: autor.

Para a realização do experimento o arquivo de treinamento foi carregado no aplicativo Weka. Empregaram-se três diferentes classificadores para avaliar o experimento, sendo os classificadores KNN, *Naive Bayes* e SVM.

5.5 Análise dos resultados

Com a realização do experimento verificou-se que o classificador que obteve melhor precisão foi o *Naive Bayes*, com 0,719, seguido pelo classificador SVM com 0,717 e KNN com 0,710. Quando se analisou o melhor classificador pelo resultado da revocação o classificador com melhor resultado foi o SVM, com 0,650, seguido pelo KNN, com 0,647, e *Naive Bayes,* 0,646. Os valores detalhados podem ser verificados no GRÁF.

Gráfico 9 - Comparação entre classificadores por precisão e revocação



Fonte: autor.

Em relação ao tempo de processamento gasto pelos classificadores, o classificar que gastou o maior tempo de processamento foi o SVM, chegando a gastar 400% mais tempo de processamento que os demais classificadores.

De acordo com a matriz confusão de cada classificador utilizado no experimento, verifica-se que, usando o classificador KNN, do total de 1.197 instâncias existentes da classe -1, 1.080 foram classificadas corretamente, o que representa 90,2% de acerto; e 117 instâncias foram classificadas incorretamente, significando 9,8% de erro. Do total de 1.328 instâncias existentes da classe 1, 554 foram classificadas corretamente, o que representa 41,7% de acerto; e 774 foram classificadas incorretamente, o que representa 58,3% de erro (FIG. 26).

Figura 26 - Resultado da matriz confusão utilizando o classificador KNN

	Instâncias Classificadas como -1	Instâncias Classificadas como 1
Instâncias da Classe -1 (Fechamento Negativo)	1080	117
Instâncias da Classe 1 (Fechamento Positivo)	774	554

Fonte: autor.

Utilizando o classificador *Naive Bayes* do total de 1.197 instâcias existentes da classe -1, 1.100 foram classificadas corretamente. o que representa 91,9% de acerto; e 97 foram classificadas, significando 8,1% de erro. Do total de 1.328 instâncias existentes da classe 1, 530 foram classificadas corretamente - 40% de acerto; e 798 foram classificadas incorretamente - 60% de erro (FIG. 27).

Figura 27 - Resultado da matriz confusão utilizando o classificador *Naive Bayes*

	Instâncias Classificadas como -1	**Instâncias Classificadas como 1**
Instâncias da Classe -1 (Fechamento Negativo)	1100	97
Instâncias da Classe 1 (Fechamento Positivo)	798	530

Fonte: autor.

Com o classificador SVM, do total de 1.197 instâncias existentes da classe -1, 1.091 foram classificadas corretamente, o que representa 91,1% de acerto; e 106 foram classificadas incorretamente, o que representa 8,9% de erro. Do total de 1.328 instâncias existentes da classe 1, 549 foram classificadas corretamente, representando 41,3% de acerto; e 779 foram classificadas incorretamente - representa 58,7% de erro, como mostra a FIG. 28.

Figura 28 - Resultado da matriz confusão utilizando o classificador SVM

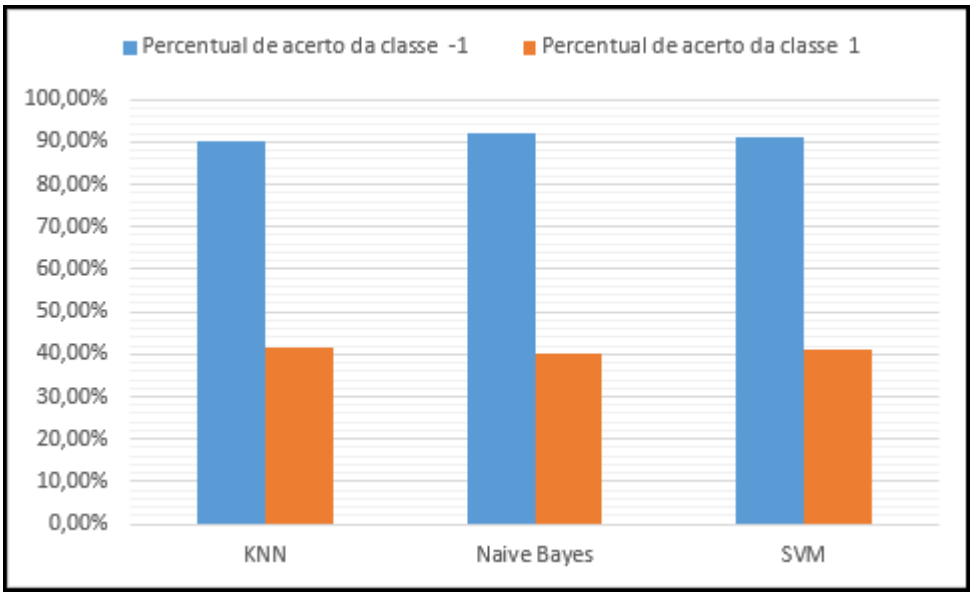
	Instâncias Classificadas como -1	**Instâncias Classificadas como 1**
Instâncias da Classe -1 (Fechamento Negativo)	1091	106
Instâncias da Classe 1 (Fechamento Positivo)	779	549

Fonte: autor.

Todos os classificadores utilizados no experimento, KNN, *Naive Bayes* e SVM, tiveram mais de 90% de acerto na classificação da classe -1, que representa

as notícias que tiveram seus sentimentos extraídos pela API *do IBM Watson Ton Analyzer* na data em que o fechamento do papel PETR4 foi negativo. O percentual de acerto da classe 1, porém, foi de 41,70% utilizando-se o classificador KNN, 40% com o *Naive Bayes* e 41,30% com o SVM. A classe 1 representa as notícias que tiveram seus sentimentos extraídos pela API *do IBM Watson Ton Analyzer* na data em que o fechamento da ação PETR4 foi positivo. No GRÁF. 10 é exibida a diferença entre o percentual de acertos entre a classe -1 e 1.

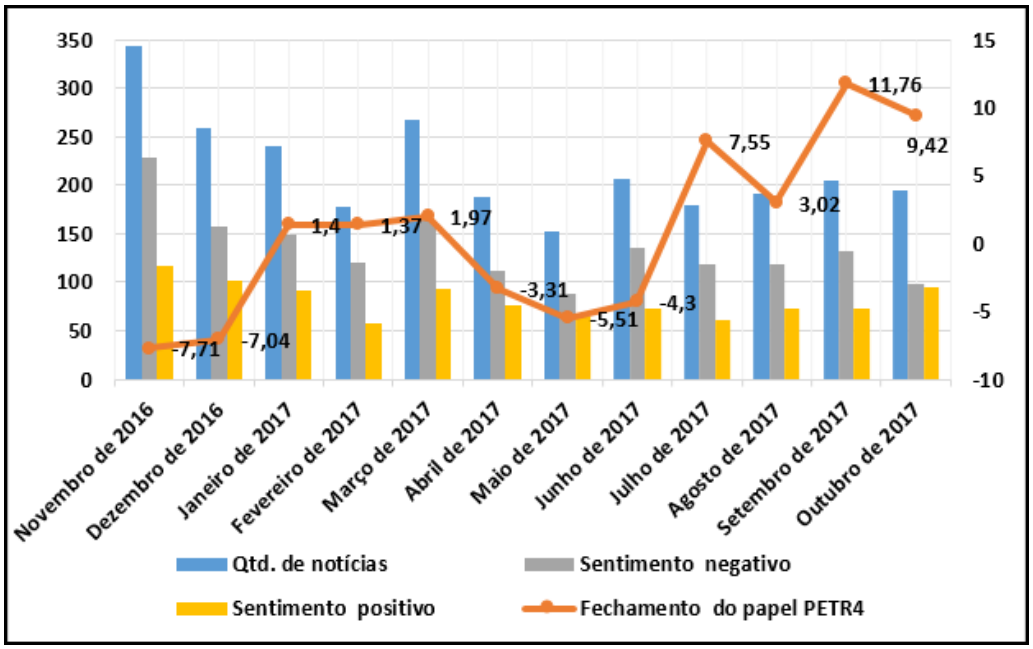
Gráfico 10 - Percentual de acertos entre a classe -1 e 1 do experimento 1



Fonte: autor.

Baseando-se nos resultados das matrizes de confusão, verifica-se que o nível de assertividade tendeu a ser superior para identificar uma tendência de baixa do que uma tendência de alta do papel PETR4 por meio dos classificadores de aprendizado de máquina testados. Fazendo uma análise somente comparando a quantidade de notícias classificadas como negativas e positivas com o fechamento mensal do papel PETR4, é possível identificar um padrão entre o fechamento negativo e a quantidade predominante de notícias negativas. O mesmo não acontece com as notícias positivas e o fechamento positivo do papel PETR4 (GRÁF. 11).

Gráfico 11 - Fechamento do papel PETR4 x quantidade de notícias classificadas como negativas e positivas no mês

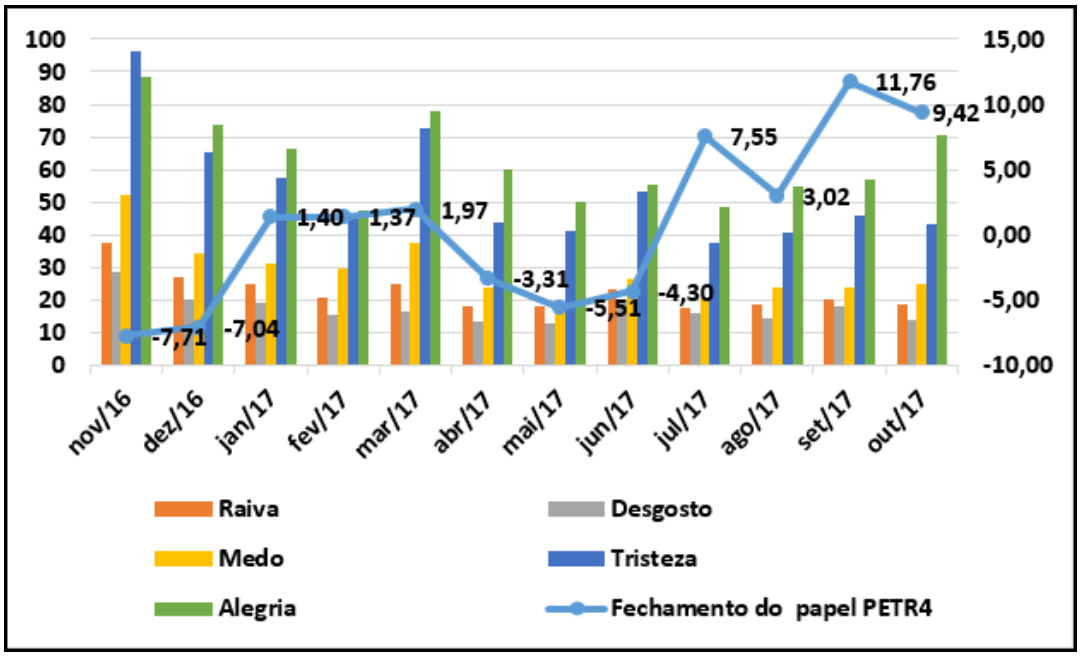


Fonte: autor.

O GRÁF. 11 mostra que, dos 12 meses analisados, em cinco o papel PETR4 fechou no negativo, sendo que nesses meses a quantidade de notícias predominante foi a de sentimento negativo e em sete meses o papel PETR4 fechou no positivo. E do total de notícias coletadas nesses meses a quantidade majoritária também foi negativa.

Analisando, porém, a soma mensal de cada índice de sentimento em vez da soma da quantidade de notícias positivas ou negativas, verifica-se que o índice predominante foi o de alegria. Entre os sentimentos negativos o predominante foi tristeza, como demonstra o GRÁF. 12.

Gráfico 12 - Soma mensal dos índices dos sentimentos e a relação com o fechamento mensal do papel PETR4



Fonte: autor.

Quando a análise foi pela soma dos índices de sentimentos em vez da quantidade de notícias positivas ou negativas verificou-se que o sentimento positivo foi superior ao negativo. Com isso se obteve um padrão diferente do anterior, em que dos 12 meses analisados em sete o papel PETR4 fechou no positivo, sendo que o sentimento predominante nesses meses foi o de alegria. E em cinco meses o papel PETR4 fechou no negativo, mas desses cinco meses um fechou com o sentimento negativo como o predominante; e nos outros quatro com o sentimento positivo.

Com esses resultados percebe-se que identificar o movimento do mercado de ações somente por meio dos sentimentos carregados pelas notícias pode ser uma tarefa complexa.

Uma explicação para os erros ocorridos no processo de classificação deve-se ao fato de que em muitas das notícias processadas, apesar de carregarem um sentimento negativo, o papel PETR4 na data em que a notícia foi publicada fechou no positivo. O mesmo aconteceu com as notícias em que, embora carregassem um sentimento positivo, o papel PETR4 na data em que a notícia foi publicada fechou no negativo. Com isso, não se pode afirmar que se em um determinado dia um conjunto

de notícias forem negativas o fechamento do dia também será negativo, pois o mercado pode interpretar as notícias de forma diferente.

Na TAB. 4, como exemplo, a notícia 01 possui sentimento negativo, cujo maior índice retornado pela API *do IBM Watson Ton Analyzer* foi o de desgosto seguido pelo de tristeza. No entanto, na data em que essa notícia foi publicada, 22/11/2016, o papel PETR4 fechou no positivo com alta de 2,25%. Uma possível causa é que, apesar de a notícia ser negativa, por tratar-se de suposto pagamento de propina, para a Petrobras foi positiva, pois agora existe uma investigação em curso e a probabilidade de os pagamentos de propinas continuarem tende a ser minimizada ao longo do tempo. Isso garante melhor resultado para a empresa a médio e longo prazo e com isso a empresa é forçada a aprimorar seus processos de governança e *compliance* (PETROBRAS1, 2018).

Outro exemplo é o da notícia 02 da TAB. 4, que possui sentimento negativo e o maior índice retornado foi de raiva seguido pelo de tristeza. Contudo, na data em que essa notícia foi publicada, 26/12/2016, o papel PETR4 fechou no positivo com alta de 1,26%. A notícia é sobre a devolução de duas concessões para exploração de blocos no estado do Mato Grosso. No primeiro momento pode parecer uma notícia negativa, mas analisando o momento pelo qual a empresa passava naquele período, com alto grau de endividamento, o conteúdo da notícia foi de encontro ao plano de desinvestimento da empresa. Com isso, mesmo com o viés negativo, foi recebida de forma positiva pelo mercado financeiro (PETROBRAS2, 2018).

Por fim, a notícia 03 publicada no dia 30/11/2016, apesar do sentimento negativo, registrou que o papel PETR4 fechou com alta de 9,14% no dia.

Tabela 3 - Notícias que foram classificadas com sentimento negativo, porém, o fechamento do papel PETR4 foi positivo no mercado de ações

Notícia 01	“Petrobras apura suposta propina envolvendo Cabral em obra do COMPERJ. em comunicado enviado à CVM, estatal se posiciona sobre notícia de que ex-governador teria recebido r$ 2,7 milhões da Andrade Gutierrez”
Índices	Raiva= 0,08; **Desgosto= 0,26**; Medo= 0,07; Tristeza= 0,20; Alegria= 0,07
Notícia 02	“Petrobras encerra atividades na Bacia do Parecis. Companhia devolveu à Agência Nacional do Petróleo duas concessões para exploração de blocos no Mato Grosso”
Índices	**Raiva= 0,63**; Desgosto= 0,15; Medo= 0,11; Tristeza= 0,32; Alegria= 0,05
Notícia 03	“Petrobras realizou 'atos temerários' para viabilizar Abreu e Lima, diz TCU. Ministro afirmou que Petrobras tinha obsessão pela construção da refinaria. TCU investiga prejuízos com a obra e vai convocar gestores.”
Índices	Raiva= 0,07; Desgosto= 0,14; Medo= 0,05; **Tristeza= 0,77**; Alegria= 0,04

Fonte: autor.

O mesmo acontece com as notícias positivas. Não se pode afirmar que se em um determinado dia um conjunto de notícias for positivo o fechamento do dia também o será. Como exemplo, na TAB. 5 na notícia 01 do dia 10/08/2017, que possui sentimento positivo, em que o maior índice retornado foi o de alegria, o preço da ação PETR4 fechou com baixa de -2,44%. Na notícia 02, publicada no dia 29/09/2017, apesar do viés positivo o papel PETR4 fechou com baixa de -0,26%. Por fim, na notícia 03, publicada no dia 03/11/2016, que também obteve índice positivo, o papel PETR4 fechou com baixa de -4,33%.

Tabela 4 - Notícias que foram classificadas com sentimento positivo, porém, o fechamento do papel PETR4 foi negativo no mercado de ações

Notícia 01	Petrobras anuncia descoberta de petróleo no pré-sal da Bacia de Campos. Segundo a Petrobras, resultado mostra “potencial de novas descobertas em baicas maduras”; óleo está em Campo de Marlim do Sul.
Índices	Raiva= 0,02; Desgosto= 0,04; Medo= 0,01; Tristeza= 0,12; **Alegria= 0,74**
Notícia 02	Aposta da EXXON no petróleo brasileiro sinaliza apetite pelo pré-sal, dizem especialistas. EXXON se uniu à Petrobras nas maiores propostas feitas no leilão; EXXON é a maior petroleira do mundo, com valor de mercado de us$ 350 bilhões.
Índices	Raiva= 0,06; Desgosto= 0,06; Medo= 0,09; Tristeza= 0,09; **Alegria= 0,68**
Notícia 03	Dinheiro devolvido por Cerveró vai todo para Petrobras, decide STF. A decisão contraria o que havia sido previsto pela Procuradoria-Geral da República (PGR) no acordo de delação premiada firmado por Cerveró
Índices	Raiva= 0,08; Desgosto= 0,06; Medo= 0,13; Tristeza= 0,11; **Alegria= 0,55**

Fonte: autor.

Avaliando de forma geral o conjunto de dados analisados no experimento por meio do arquivo de treinamento executado na ferramenta WEKA, a precisão

encontrada foi superior à de outros trabalhos realizados na mesma linha de pesquisa, podendo-se citar os trabalhos de Schumaker *et al.* (2012), Kim, Jeong e Ghani (2014), Junqué de Fortuny *et al.* (2014), Li, Q. *et al.* (2014), Gunduz e Cataltepe (2015), Ichinose e Shimada (2016) e Tirea e Negru (2016), que obtiveram precisão abaixo de 70%.

O desenvolvimento e resultado desta pesquisa ficaram próximos do trabalho de Duong, Nguyen e Dang (2016), que buscou identificar tendências do mercado de ações do Vietnã, especificamente do índice VN30, por meio de notícias financeiras.

Para efeitos de comparação entre os dois trabalhos, na pesquisa de Duong, Nguyen e Dang (2016) utilizou-se um *corpus* com 1.884 notícias. Neste trabalho foi empregado um *corpus* com 2.525 notícias. O período de coleta no trabalho de Duong, Nguyen e Dang (2016) foi de um ano - 1º/05/2014 a 30/04/2015. Este trabalho também coletou notícias de 1º/11/2016 a 31/10/2017. A forma de validação de ambos os experimentos foi a mesma, com a avaliação por meio das matrizes de confusão, precisão, revocação. A forma adotada para o treinamento diferiu pelo fato de Duong, Nguyen e Dang (2016) utilizarem eu seu trabalho 70% (1.319) do *corpus* para treinamento e 30% (565) para teste. Já nesta pesquisa adotou-se como forma de treinamento a validação cruzada (K-Fold).

Em relação aos classificadores empregados no trabalho de Duong, Nguyen e Dang (2016), foi escolhido somente o SVM. Para esta pesquisa usou-se, além do classificado SVM, o KNN e *Naive Bayes*. A precisão obtida em ambas as pesquisas foi próxima do trabalho de Duong, Nguyen e Dang (2016), que destacou precisão de 68% utilizando o *corpus* referente a um ano de coleta. Esta pesquisa alcançou a precisão de 72%. Apesar dos 72% de precisão na identificação de tendências alcançados no experimento, percebe-se que a abordagem para utilizar os sentimentos carregados pelas notícias enfrenta uma série de desafios, como compreender os fundamentos e a variação do impacto que essas notícias exercem sobre os investidores ao longo do tempo, além de determinar quais notícias específicas atraem especuladores ou têm impacto limitado.

5.6 Comunicação e contribuições gerais desta pesquisa

Hevner *et al.* (2004) indicam a necessidade de comunicação para difundir o conhecimento resultante de uma pesquisa. A comunicação desta pesquisa se deu a partir do seu processo de formalização.

Como síntese das contribuições realizadas por esta pesquisa podem-se citar os artefatos gerados que podem ser utilizados por outros pesquisadores, como, por exemplo, para encontrar novas formas e técnicas de trabalhar como o *corpus* de notícias disponibilizado, como também avaliar a utilização de outros classificadores diferentes dos testados nesta pesquisa. No geral, este estudo está disponibilizando os seguintes artefatos:

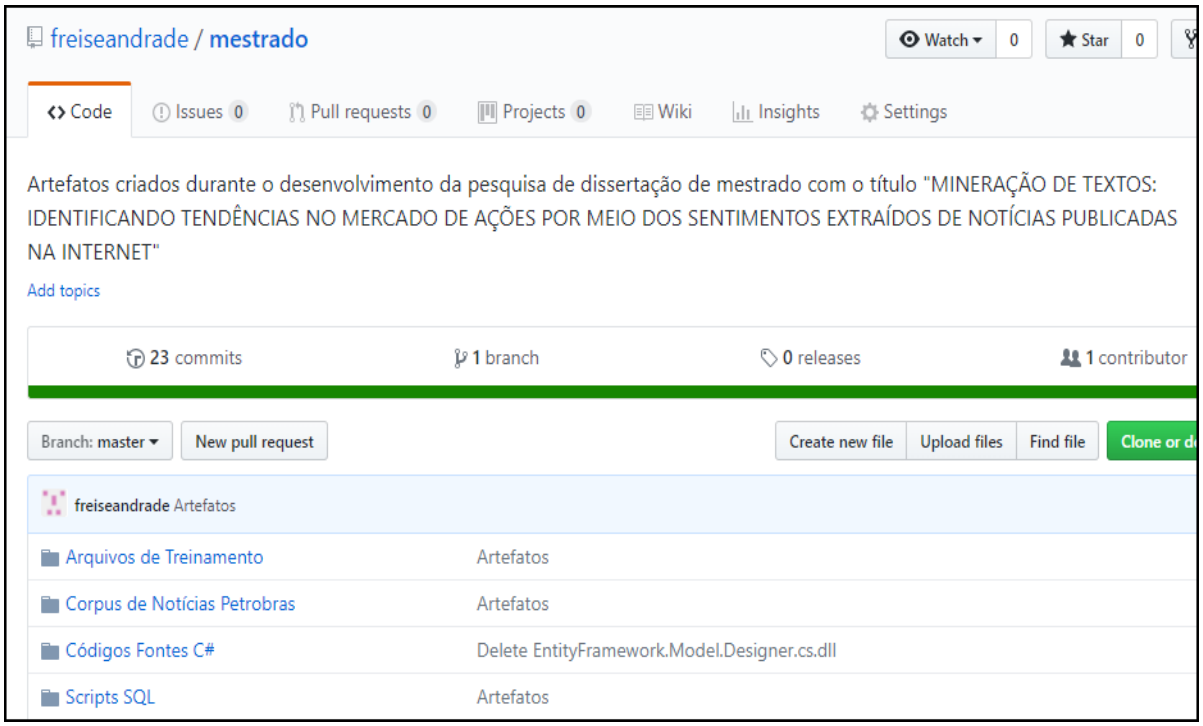
a) Um *corpus* de notícias sobre a Petrobras disponibilizadas no formato de texto coletadas no período de 1º/11/2016 a 01/11/2017;
b) um *corpus* com o fechamento diário do papel PETR4 no período de 1º/11/2016 a 1º/11/2017;
c) algoritmos computacionais desenvolvidos em C# para extrair os *feeds* de notícias e criação dos arquivos de treinamentos no formato exigido pelo WEKA;
d) arquivo de treinamento utilizado nesta pesquisa descrito na seção 5.3.7.1;
e) *scripts* SQL para a criação das tabelas e *views* utilizadas nesta pesquisa
f) o desenvolvimento de uma revisão sistemática da literatura enumerando as técnicas mais utilizadas para identificação de tendências no mercado de ações por meio de notícias publicadas na internet;

Esta pesquisa também contribui para a geração de conhecimento com base na investigação, avaliação e validações realizadas por meio da extração e utilização dos sentimentos extraídos das notícias para identificar tendências no mercado de ações.

Os artefatos gerados no desenvolvimento desta pesquisa foram disponibilizados e podem ser acessados no GitHub (GitHub.com), um repositório de código de hospedagem baseado no sistema de controle de versão Git. O GitHub é um exemplo de um espaço de trabalho baseado no conhecimento. Esse *site* integra

uma série de recursos sociais que tornam visível a informação única sobre usuários e suas atividades dentro e entre projetos de *software* de código aberto (DABBISH *et al.*, 2012). O endereço para acessar os artefatos desenvolvidos durante a pesquisa é https://github.com/freiseandrade/mestrado, que deve exibir uma página como a da FIG. 29.

Figura 29 - Tela inicial para acesso aos artefatos desenvolvidos



Fonte: autor.

6 CONSIDERAÇÕES FINAIS

O presente estudo objetivou avaliar a precisão de um sistema de informação para identificar a tendência do papel PETR4 da empresa Petrobras na BM&F Bovespa por meio da extração dos sentimentos dos *feeds* de notícias publicados na internet. Com base nos resultados obtidos no desenvolvimento desta pesquisa, verificou-se que é possível medir o sentimento expressado por meio das notícias publicadas na internet e que estas podem influenciar as ações dos investidores e, consequentemente, podem exercer efeitos importantes sobre o mercado de ações.

Para atingir o objetivo proposto, os seguintes objetivos específicos foram estabelecidos: descrever métodos e técnicas comumente empregados na identificação de tendências no mercado de ações por meio de notícias publicadas na internet; construir um *corpus* de notícias financeiras sobre a Petrobras; elaborar algoritmos computacionais para extração das notícias e criação dos arquivos de treinamento. Vale ressaltar que tanto o objetivo geral quanto os objetivos específicos foram alcançados por esta pesquisa por meio da revisão sistemática da literatura e da metodologia *Design Science Research* adotada no seu desenvolvimento.

Com o *corpus* de notícias criado para a realização dos experimentos alcançou-se precisão de 72% na identificação de tendências do papel PETR4.

Ambos os algoritmos, *Naive Bayes* e SVM, apresentaram desempenho superior ao KNN. Quando se analisa de forma detalhada o retorno da matriz confusão de cada classificador, verifica-se que o nível de assertividade da classe -1 foi de mais de 90% de acerto na classificação, o que significa que, para o classificador KNN, do total de 1.197 instâncias da classe -1, 1.080 foram classificadas corretamente. No classificador *Naive Bayes,* do total de 1.197 instâncias da classe -1, 1.100 foram classificadas corretamente. E para o classificador SVM, do total de 1.197 instâncias existentes da classe -1, 1.091 foram classificadas corretamente. A classe -1 representa as notícias que tiveram seus sentimentos extraídos pela API da *IBM Watson Ton Analyzer* na data em que o fechamento da ação PETR4 foi negativo.

O percentual de acerto da classe 1 foi de somente 41,70% utilizando o classificador KNN, o que significa que, do total de 1.328 instâncias existentes da classe 1, 554 foram classificadas corretamente; 40% utilizando o *Naive Bayes* com

530 instâncias classificadas corretamente; e 41,30% utilizando o SVM com 549 instâncias classificadas corretamente. A classe 1 representa as notícias que tiveram seus sentimentos extraídos pela API do IBM *Watson Ton Analyzer* na data em que o fechamento da ação PETR4 foi positivo. Com esses resultados verifica-se que o nível de assertividade tende a ser superior na identificação de tendências de baixa do papel PETR4 comparado com a identificação das tendências de alta. Os erros de classificação ocorrem muitas vezes pelo fato de que muitas das notícias processadas, apesar de carregarem um sentimento negativo, foram publicadas na data em que o papel PETR4 fechou em alta. O mesmo aconteceu com as notícias que, embora carregassem um sentimento positivo, na data em que foram publicadas o papel PETR4 fechou em baixa.

Analisando esses resultados, percebe-se que mesmo diante de 72% de precisão obtidos no decorrer dos experimentos, a abordagem para utilizar os sentimentos carregados pelas notícias enfrenta uma série de desafios, como compreender os fundamentos e a variação do impacto que essas notícias exercem sobre os investidores ao longo do tempo, além de determinar quais notícias específicas atraem especuladores ou têm impacto limitado.

Com isso, concluiu-se que ainda há muito a ser feito em termos de explicação desse quadro, mas os potenciais retornos de melhor compreensão do sentimento dos investidores são substanciais para o aumento da precisão na identificação da tendência de uma ação. E as notícias publicadas podem ser importantes meios para atingir esse objetivo.

6.1 Limitação da pesquisa e trabalhos futuros

O mercado de ações dentro de sua dinâmica pensa no valor futuro e não somente nas notícias publicadas do dia, pois elas são importantes caso impactem o desempenho da empresa no futuro. Além disso, existem muitas variáveis que podem afetar o resultado da pesquisa, como, por exemplo, a entrada e saída de dólares no país ou um acontecimento internacional de grandes proporções. Tem-se também a temporalidade da informação, relativa ao passado e futuro. Tem-se a limitação da tradução dos textos de notícias para o inglês, já que a API utilizada para a extração dos sentimentos não contém o idioma português.

Para trabalhos futuros, sugere-se aprimorar o sistema de identificação de tendências, ampliando o *corpus* a ser utilizado. Com essa ampliação, além dos *feeds* de notícias relevantes sobre determinada empresa, deve-se passar a utilizar notícias sobre a economia nacional e internacional e adotar também fontes adicionais de dados como *posts* de redes sociais. E também passar a considerar a temporalidade desses dados.

REFERÊNCIAS

ABDULHAMID, M.S.M.; OSHO, O.; ISMAILA, I. **Comparative analysis of classification algorithms for Email spam detection**. December 2017, 2018. Disponível em: <https://www.researchgate.net/profile/Shafii_Abdulhamid2/publication/322129882_Comparative_Analysis_of_Classification_Algorithms_for_Email_Spam_Detection/links/5a4670e0458515f6b0558b0a/Comparative-Analysis-of-Classification-Algorithms-for-Email-Spam-Detect>. Acesso em: maio de 2018.

ABDULLAH, S.S.; RAHAMAN, M.S.; RAHMAN, M.S. Analysis of stock market using text mining and natural language processing. **2013 International Conference on Informatics, Electronics and Vision (ICIEV)**, p. 1–6, 2013. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6572673>. Acesso: maio de 2018.

ACKERT, L.F.; CHURCH, B.K.; DEAVES, R. Emotion and financial markets. **Economic Review**, v. 88, n. 2, p. 33, 2003.

AGARWAL, S.; SUREKA, A. Spider and the flies : Focused Crawling on Tumblr to Detect Hate Promoting Communities. **European Intelligence and Security Informatics Conference**, p. 9164, 2016. Disponível em: <http://arxiv.org/abs/1603.09164>. Acesso em: maio de 2018.

ALOSTAD, H.; DAVULCU, H. Directional prediction of stock prices using breaking news on twitter. **2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)**, 2015.

ARMBRUST, M. *et al*. A view of cloud computing clearing the clouds away from the true potential and obstacles posed by this computing capability. **Communications of the ACM**, v. 53, n. 4, p. 50–58, 2010.

ARPACI, U.; KARAOGLU, S.; AYVAZ, S. A deep learning approach for optimization of systematic signal detection in financial trading systems with big data. **International Journal of Intelligent Systems and Applications in Engineering**, v. SpecialIssue, n. SpecialIssue, p. 31–36, 2017. Disponível em: <https://www.atscience.org/IJISAE/article/view/591/pdf>. Acesso em: 14/1/2018.

ATTIGERI, G.V. *et al.* Stock market prediction: A big data approach. **35th IEEE Region 10 Conference, TENCON 2015**, v. 2016–January, 2016. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84962185156&partnerID=40&md5=c47542775423d3234efde26bf948ed9d>. Acesso em: abril de 2018.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de informação: conceitos e tecnologia das máquinas de busca**. 2. ed., 2013.

BATES, M. Models of natural language understanding. **Proceedings of the National Academy of Sciences**, v. 92, n. 22, p. 9977–9982, 1995. Disponível em: <http://www.pnas.org/content/92/22/9977.abstract>. Acesso em: janeiro de 2018.

BHARDWAJ, A. *et al*. Sentiment analysis for indian stock market prediction using sensex and nifty**. Procedia Computer Science**, v. 70, p. 85–91, 2015.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of Computational Science**, v. 2, n. 1, p. 1–8, 2011.

BOTTA, A. *et al.* Integration of cloud computing and internet of things: A SURVEY. **Future Generation Computer Systems**, v. 56, p. 684-700, 2016.

BRANCH, W.A.; EVANS, G.W. Learning about risk and return: A simple model of bubbles and crashes. **American Economic Journal: Macroeconomics American**, v. 3, n. 3, p. 159-191, 2011. Disponível em: <http://www.jstor.org/stable/41237157%0Ahttp://about.jstor.org/terms%0Ahttp://www.aeaweb.org/articles.php?doi=10.>. Acesso em: fevereiro de 2018.

BRITO, E.M.N. **Mineração de textos**: detecção automática de sentimentos em comentários nas mídias sociais. Dissertação (Mestrado em Sistemas de Informação) - FUMEC, p. 54, 2016.

CARBONELL, J.G.; MICHALSKI, R.S.; MITCHELL, T.M. An overview of machine learning. *In*: MICHALSKI, R.S.; CARBONELL, J.G.; MITCHELL, T.M. (Orgs.). **Machine learning: An Artificial Intelligence Approach,** p.3-23, 1983. Berlin, Heidelberg: Springer Berlin Heidelberg. Disponível em: <https://doi.org/10.1007/978-3-662-12405-5_1>. Acesso em: março de 2018.

CHOWDHURY, G.G. Natural language processing. **Annual Review of Information Science and Technology**, v. 37, n. 1, p. 51-89, 2005. Wiley Subscription Services, Inc., A Wiley Company. Disponível em: <http://doi.wiley.com/10.1002/aris.1440370103>. Acesso em: 11/1/2017.

DABBISH, L. *et al.* Social coding in GitHub: transparency and collaboration in an open software repository. **Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work**, p. 1277-1286, 2012. Disponível em: <http://dl.acm.org/citation.cfm?doid=2145204.2145396%5Cnhttp://dl.acm.org/citation.cfm?id=2145204.2145396%5Cnhttp://dl.acm.org/citation.cfm?doid=2145204.2145396>. Acesso em: junho de 2018.

DENG, X. *et al.* An improved method to construct basic probability assignment based on the confusion matrix for classification problem. **Information Sciences**, 2016. Elsevier Inc. Disponível em: <http://dx.doi.org/10.1016/j.ins.2016.01.033>. Acesso em: dezembro de 2017.

DIESTE, O.; GRIMÁN, A.; JURISTO, N. Developing search strategies for detecting relevant experiments. **Empirical Software Engineering**, v. 14, n. 5, p. 513–539, 2009. Disponível em: <https://doi.org/10.1007/s10664-008-9091-7>. Acesso em: junho de 2018.

DANG, M.; DUONG, D. Improvement methods for stock market prediction using financial news articles. **National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)**, v.3 ,p. 125–129, 2016.

DRESCH, A.; LACERDA, D.P.; ANTUNES JÚNIOR, J.A.V. Design Science Research: método de pesquisa para avanço da ciência e tecnolgia. **Gestão Produção**, v. 20, n. 4, p. 741–761, 2013.

DUONG, D.; NGUYEN, T.; DANG, M. Stock market prediction using financial news articles on ho chi minh stock exchange. *In*: 10th INTERNATIONAL CONFERENCE ON UBIQUITOUS INFORMATION MANAGEMENT AND COMMUNICATION. **Proceedings...** , IMCOM '16. p.71:1--71:6, 2016. New York, NY, USA: ACM. Disponível em: <http://doi.acm.org/10.1145/2857546.2857619>. Acesso em: junho de 2018.

EHRENTREICH, N. Technical trading in the Santa Fe Institute Artificial Stock Market revisited. **Journal of Economic Behavior and Organization**, v. 61, n. 4, p. 599-616, 2006.

ECMA, INTERNATIONAL. **The JSON Data Interchange Syntax**. 2. edition. [S.l.: s.n.], 2017. 16 p.

FAN, W.; WATANABE, T. Dynamic prediction of forthcoming trends in stock prices from news articles. 2Nd INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, MINING AND SEMANTICS. **Proceedings...**, WIMS '12. p.16:1-16:9, 2012. New York, NY, USA: ACM. Disponível em: <http://doi.acm.org/10.1145/2254129.2254151>. Acesso em: dezembro de 2017.

ECONOMATICA. **As ações mais rentáveis com 100% de presença nos últimos 10 anos**. Disponível em: <https://economatica.com/estudos/data/20180219a.pdf>. Acesso em: 29 jun. 2018.

ESKICI, H.B.; KOÇAK, N.A. A text mining application on monthly price developments reports. **Central Bank Review**, v. 18, n. 2, p. 51-60, jun. 2018.

FEIN, M.L. Robo-advisors: A closer look. **SSRN Electronic Journal**, 2015. Disponível em: <http://www.ssrn.com/abstract=2658701>. Acesso em: 14/1/2018.

FELDMAN, R.; SANGER, J. **The text mining handbook:** advanced approaches in analyzing unstructured data. Cambridge University Press, 2006.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook**. Cambridge University Press, 2007.

FIELDING, R.T.; TAYLOR, R.N. Principled design of the modern Web architecture. **ACM Transactions on Internet Technology**, v. 2, n. 2, p. 115=150, 2002. Disponível em: <http://portal.acm.org/citation.cfm?doid=514183.514185>. Acesso em: fevereiro de 2018.

FLACH, P.; KULL, M. Precision-recall-gain curves: PR analysis done right. **Advances in Neural Information Processing Systems 28**, v. 1, p. 838-846, 2015. Disponível em: <http://papers.nips.cc/paper/5867-precision-recall-gain-curves-pr-analysis-done-right.pdf>. Acesso em: fevereiro de 2018.

GEVA, T.; ZAHAVI, J. Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. **Decision Support Systems**, v. 57, p. 212-223, 2014. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0167923613002418>. Acesso em: 9/3/2017.

GIDOFALVI, G. **Using news articles to predict stock price movements**. 2001. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/similar;jsessionid=D7D708D1BCE52DED25C429CC95900765?doi=10.1.1.17.5103&type=ab>. Acesso em: março de 2018.

GUNDUZ, H.; CATALTEPE, Z. Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. **Expert Systems with Applications**, v. 42, n. 22, p. 9001-9011, 2015.

GUNTHER, M. **Os axiomas de Zurique**. Best Business, 4. ed., 2017.

HAGENAU, M.; LIEBMANN, M.; NEUMANN, D. Automated news reading: Stock price prediction based on financial news using context-capturing features. **Decision Support Systems**, v. 55, n. 3, p. 685-697, 2013. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.dss.2013.02.006>. Acesso em: março de 2018.

HARVARD LAW. **RSS 2.0 Specification. 2003**. Disponível em: <https://goo.gl/ejDAZS>. Acesso em: 12 set. 2017.

HEVNER, A.R. *et al.* Design science in information systems research. **MIS Quarterly**, v. 28, n. 1, p. 75-105, 2004. Disponível em: <http://dblp.uni-trier.de/rec/bibtex/journals/misq/HevnerMPR04>. Acesso em: janeiro de 2018.

HILLIER, D.; GRINBLATT, M.; TITMAN, S. **Financial markets and corporate strategy**. McGraw Hill, 2011.

HU, M.; LIU, B. Mining and summarizing customer reviews. *In:* 2004 ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MININg - KDD '04. **Proceedings...,** p. 168, 2004. New York, USA: ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=1014052.1014073>. Acesso em: 13/8/2017.

IBM1, **Watson tone analyzer**: introduction. 2018a. Disponível em: <https://www.ibm.com/watson/developercloud/tone-analyzer/api/v3/#introduction>. Acesso em: 03 jan. 2018.

IBM2, **Watson language translator:** introduction. 2018b. Disponível em: <https://www.ibm.com/watson/developercloud/language-translator/api/v2/#introduction>. Acesso em: 03 jan. 2018.

ICHINOSE, K.; SHIMADA, K. Stock market prediction from news on the web and a new evaluation approach in trading. *In***:** 2016 5th IIAI INTERNATIONAL CONGRESS ON ADVANCED APPLIED INFORMATICS, IIAI-AAI 2016, p. 77-81. **Proceedings…,** 2016.

INGLE, V.; DESHMUKH, S. Hidden markov model implementation for prediction of stock prices with TF-IDF Features. *In:* INTERNATIONAL CONFERENCE ON ADVANCES IN INFORMATION COMMUNICATION TECHNOLOGY & COMPUTING**. Proceedings**...: AICTC '16.New York, NY, USA: ACM, 2016.

JELIAZKOVA, N.; JELIAZKOV, V. AMBIT RESTful web services: An implementation of the OPENTOX application programming interface. **Journal of Cheminformatics**, v. 3, n. 1, p. 1-18, 2011.

JOHNSON, E. *et al.* Affective avatar interactions: towards recognizing emotions in verbal interaction. **Cham**, Springer, p. 05–310, 2017. Disponível em: <http://link.springer.com/10.1007/978-3-319-67585-5_32>. Acesso em: 9/5/2018.

JUNQUÉ DE FORTUNY, E. *et al.* Evaluating and understanding text-based stock price prediction models. **Information Processing and Management**, v. 50, n. 2, p. 426-441, 2014.

KAMINSKI, J. T**he science behind the service**. Disponível em: <https://github.com/IBM-Bluemix-Docs/tone-analyzer/blob/master/science.md>. Acesso em: 10 maio 2018.

KHADJEH NASSIRTOUSSI, A. *et al.* Text mining for market prediction: A systematic review. **Expert Systems with Applications**, v. 41, n. 16, p. 7653-7670, 2014.

KHADJEH NASSIRTOUSSI, A. *et al.* Text mining of news-headlines for FOREX market prediction: A Multi-layer dimension reduction algorithm with semantics and sentiment. **Expert Systems with Applications**, v. 42, n. 1, p. 306-324, 2015. Elsevier Ltd. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2014.08.004>. Acesso em: maio de 2018.

KIM, Y.; JEONG, S.R.; GHANI, I. Text opinion mining to analyze news for stock market prediction. **International Journal of Advances in Soft Computing and Its Applications**, v. 6, n. 1, p. 1-13, 2014.

KIRILENKO, A.A.; LO, A.W. Moore's law vs. Murphy's law: algorithmic trading and its discontents. **SSRN Electronic Journal**, 2013. Disponível em: <http://www.ssrn.com/abstract=2235963>. Acesso em: 3/5/2018.

KITCHENHAM, B. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, p. 1-26, 2004.

KLOPTCHENKO, A. *et al*. Combining data and text mining techniques for analysing financial reports. **Intelligent Systems in Accounting, Finance and Management**, v. 12, n. 1, p. 29-41, 2004.

LI IM, T. *et al*. Analysing market sentiment in financial news using lexical approach. *In*: IEEE CONFERENCE ON OPEN SYSTEMS, ICOS 2013,**Anais…,** p. 145–149, 2013.

LI, Q. *et al*. The effect of news and public mood on stock movements. **Information Sciences**, v. 278, p. 826-840, 2014.

LI, Q. *et al*. A tensor-based information framework for predicting the stock market. **ACM Trans. Inf. Syst.**, v. 34, n. 2, p. 11:1–11:30, fev. 2016.

LI, X. *et al.* News impact on stock price return via sentiment analysis. **Knowledge-Based Systems**, v. 69, n. 1, p. 14-23, 2014. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.knosys.2014.04.022>. Acesso em: junho de 2018.

LIAM, L. **Extensible markup language (XML)**. Disponível em: <https://www.w3.org/XML/>. Acesso em: 10 maio 2018.

LIU, B. **Sentiment analysis and opinion mining**. Morgan & Claypool, 2012.

LOPES, T.J.P. *et al.* Mineração de opiniões aplicada à análise de investimentos. *In:* COMPANION PROCEEDINGS OF THE XIV BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB. **Anais...**, WebMedia ’08. p.117-120, 2008. New York, NY, USA: ACM. Disponível em: <http://doi.acm.org/10.1145/1809980.1810012>. Acesso em: março de 2018.

MAO, H.; COUNTS, S.; BOLLEN, J. Predicting financial markets: comparing survey, news, twitter and search engine data. **arXiv Preprint**, p. 10, 2011. Disponível em: <http://arxiv.org/abs/1112.1051>. Acesso em: março de 2018.

MARTINAZZO, B.; PARAISO, E.C. Identificação de emoções em notícias curtas. **CLEI-Conferência Latino-Americana de Informática**, v. 1, p. 1–10, 2010. Disponível em: <http://www.ppgia.pucpr.br/~paraiso/mineracaodeemocoes/recursos/219_CLEI_Martinazzo_Paper.pdf>. Acesso em: fevereiro de 2018.

METE, M. *et al.* Knowledge discovery in textual databases: a concept-association mining approach. *In*: CHAN, Y; TALBURT, J.; TALLEY, T.M. (Orgs.); **Data engineering: mining, information and intelligence**. p.225-243, 2010. Boston, MA: Springer US. Disponível em: <https://doi.org/10.1007/978-1-4419-0176-7_11>. Acesso em: junho de 2018.

MICROSOFT1, Corporation. VISUAL Studio Community 2017. V**ersion 14. Microsoft Corporation, 2017**. Disponível em: <https://goo.gl/jU2Urg>. Acesso em: 16 set. 2017.

MICROSOFT2, Corporation. SQL Server 2016 SP1 Express Edition. V**ersion 14. Microsoft Corporation, 2017**. Disponível em: <https://goo.gl/zQtr2p>. Acesso em: 16 set. 2017.

MICHELLC, D. *et al.* Innovación para el análisis de sentimientos en texto, una revisión de la técnica actual aplicando metodologías de crowdsourcing Innovation for the analysis of feelings in text, a revision of the current technique applying crowdsourcing methodologies. *In:* **I**X CONGRESO INTERNACIONAL DE LA RED DE INVESTIGACIÓN Y DOCENCIA EN INNOVACIÓN TECNOLÓGICA. **Anais**..., 2017.

MOSTAFA, M. *et al.* **Incorporating emotion and personality-based analysis in user-centered modelling**. 2016. Disponível em: <http://arxiv.org/abs/1608.03061>. Acesso em: abril de 2018.

MUKUND, Y.R. *et al*. Influence of news on individual confidence bias in stock markets. *In:* THE 11th INTERNATIONAL KNOWLEDGE MANAGEMENT IN ORGANIZATIONS CONFERENCE ON THE CHANGING FACE OF KNOWLEDGE MANAGEMENT IMPACTING SOCIETY. **Proceedings...**, KMO '16. p.20:1-20:9, 2016. New York, NY, USA: ACM. Disponível em: <http://doi.acm.org/10.1145/2925995.2926001>. Acesso em: novembro de 2017.

NASUKAWA, T.; NAGANO, T. Text analysis and knowledge mining system. **IBM Systems Journal**, v. 40, n. 4, p. 967-984, 2001.

NIGRO, T. **Robôs investidores na bolsa de valores**. Disponível em: <http://oprimorico.com.br/o-caminho/inteligencia-artificial-mais-rica-do-que-voce-imagina/>. Acesso em: 08 jan. 2018.

NIZER, P.S.M.; NIEVOLA, J.C. Predicting published news effect in the Brazilian stock market. **Expert Systems with Applications**, v. 39, n. 12, p. 10674-10680, 2012.

NOFSINGER, J.R. Social mood and financial economics. **Journal of Behavioral Finance**, v. 6, n. 3, p. 144-160, 2005. Disponível em: <http://www.tandfonline.com/doi/abs/10.1207/s15427579jpfm0603_4>. Acesso em: 12/11/2016.

OLIVEIRA, N.; CORTEZ, P.; AREAL, N. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from twitter. *In:* 3rd INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, MINING AND SEMANTICS. **Proceedings...** , WIMS '13. p.31:1-31:8, 2013. New York, NY, USA: ACM. Disponível em: <http://doi.acm.org/10.1145/2479787.2479811>. Acesso em: outubro de 2017.

ONG, S.P. *et al*. The materials application programming interface (API): A simple, flexible and efficient API for materials data based on REpresentational state transfer (REST) principles. **Computational Materials Science**, v. 97, p. 209-215, 2015. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.commatsci.2014.10.037>. Acesso em: março de 2018.

PARANÁ, E. A digitalização do mercado de capitais no Brasil: tendências recentes. **Boletim de Economia e Política Internacional**, v. 23, 2017.

PEFFERS, K. *et al.* A design science research methodology for information systems research. **Journal of Management Information Systems**, v. 24, n. 3, p. 45–77, 2007. Disponível em: <http://www.tandfonline.com/doi/abs/10.2753/MIS0742-1222240302>. Acesso em: junho de 2018.

PETROBRAS1. **Focamos na recuperação financeira**. Disponível em: <https://seguindoemfrente.hotsitespetrobras.com.br/index.htm#focamos-na-recuperacao-financeira>. Acesso em: 26 mar. 2018.

PETROBRAS2. **Nos esforçamos em aprimorar processos de governança e compliance**. Disponível em: <https://seguindoemfrente.hotsitespetrobras.com.br/index.htm#nos-esforcamos-em-aprimorar-processos-de-governanca-e-compliance>. Acesso em: 26 mar. 2018.

PETROBRAS3. **Quem somos**. Disponível em: < ttp://www.petrobras.com.br/pt/quem-somos/perfil/>. Acesso em: 20 mai. 2017. <http://www.bmfbovespa.com.br/pt_br/servicos/negociacao/data-center/data-center/co-location-em-spa.htm/>. Acesso em: 03 mai. 2018.

PICHILIANI, M. **Usando views.** Disponível em: <https://imasters.com.br/artigo/239/sql-server/usando-views?trace=1519021197&source=single>. Acesso em: 03 jan. 2018.

PORTNOY, K. High frequency trading and the stock market: A look at the effects of Trade volume on stock price changes. **The Park Place Economist**, v. 19, n. 1, 2011.

PRÖLLOCHS, N.; FEUERRIEGEL, S.; NEUMANN, D. Enhancing sentiment analysis of financial news by detecting negation scopes. *In:* ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES. **Proceedings…**, v. 2015–March, p. 959-968, 2015.

RAHMAN, I. *et al*. Ergothioneine inhibits oxidative stress- and TNF-alpha-induced NF-kappa B activation and interleukin-8 release in alveolar epithelial cells. **Biochemical and biophysical research communications**, v. 302, n. 4, p. 860-4, 2003. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/12646250>. Acesso em: dezembro de 2017.

SAITO, R.; TULIO, M.; PADILHA, C. Por que as empresas fecham o capital no Brasil ? **Revista Brasileira de Finanças** (Online), Rio de Janeiro, v. 13, n. 2, p. 1-35, 2013.

SANTOS, J.O.; SANTOS, J.A.R. Mercado de capitais: racionalidade versus emoção. **Revista Contabilidade & Finanças**, v. 16, p. 103-110, 2005. scielo. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1519-70772005000100008&nrm=iso>. Acesso em: fevereiro de 2018.

SCHUMAKER, R.P. *et al.* Evaluating sentiment in financial news articles. **Decision Support Systems**, v. 53, n. 3, p. 458-464, 2012. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.dss.2012.03.001>. Acesso em: novembro de 2017.

SEBASTIANI, F. Machine learning in automated text categorization. **Association for Computing Machinery - ACM Computing Surveys**, v. 34, n. 1, p. 1-47, 2002. New York, NY, USA: ACM. Disponível em: <http://doi.acm.org/10.1145/505282.505283>. Acesso em: janeiro de 2018.

SINGH, G.; SAHU, S. Review on “really simple syndication (RSS) technology tools”. *In:* **2**015 IEEE INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE AND COMMUNICATION TECHNOLOGY, CICT. **Proceedings…,** 2015, p. 757-761, 2015.

SRINIVASAN, S. **Cloud computing basics**. 2014. Disponível em: <http://link.springer.com/10.1007/978-1-4614-7699-3>. Acesso em: novembro de 2017.

SUHADOLNIK, N.; GALIMBERTI, J.; DA SILVA, S. Robot traders can prevent extreme events in complex stock markets. **Physica A: Statistical Mechanics and its Applications**, v. 389, n. 22, p. 5182–5192, 2010. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.physa.2010.07.025>. Acesso em: junho de 2018.

TABOADA, M. *et al.* Lexicon-based methods for sentiment analysis. **Computational Linguistics**, v. 37, n. 2, p. 267-307, 2011. MIT Press. Disponível em: <http://www.mitpressjournals.org/doi/10.1162/COLI_a_00049>. Acesso em: 15/1/2017.

TESO, E. *et al*. Application of text mining techniques to the analysis of discourse in eWOM communications from a gender perspective. **Technological Forecasting and Social Change**, v. 129, p. 131-142, April 2018.

TIREA, M.; NEGRU, V. Text mining news system: Quantifying certain phenomena effect on the stock market behavior. *In:* 17th INTERNATIONAL SYMPOSIUM ON SYMBOLIC AND NUMERIC ALGORITHMS FOR SCIENTIFIC COMPUTING, SYNASC 2015. **Proceedings…**, p. 391-398, 2016.

WAIKATO, University. Weka. **Version 3.** University Of Waikato, 2017. Disponível em: <https://goo.gl/M5zVXJ>. Acesso em: 16 set. 2017.

w3schools, XML RSS. Disponível em<https://www.w3schools.com/xml/xml_rss.asp>. Acesso em: 03 mai. 2018.

YADAV, S.; SHUKLA, S. Analysis of k-Fold cross-validation over hold-out validation on colossal datasets for quality classification. *In*: 6TH INTERNATIONAL ADVANCED COMPUTING CONFERENCE, IACC 2016. **Proceedings…**, n. Cv, p. 78-83, 2016.

YE, M.; YAO, C.; GAI, J. The externality of high frequency trading. **SSRN Electronic Journal**, 2012. Disponível em: <http://www.ssrn.com/abstract=2066839>. Acesso em: 3/5/2018.

WEI, Y.-C. *et al.* Informativeness of the market news sentiment in the Taiwan stock market. **The North American Journal of Economics and Finance**, v. 39, p. 158-181, 2017.