Juvenal José Duarte

Influência dos textos de notícias na queda de preços no mercado de ações brasileiro

Sorocaba, SP

2019

Juvenal José Duarte

Influência dos textos de notícias na queda de preços no mercado de ações brasileiro

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCC-So) da Universidade Federal de São Carlos como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação. Linha de pesquisa: Inteligência Artificial e Banco de-Dados.

Universidade Federal de São Carlos – UFSCar

Centro de Ciências em Gestão e Tecnologia – CCGT

Programa de Pós-Graduação em Ciência da Computação – PPGCC-So

Orientador: Profa. Dra. Sahudy Montenegro González

Sorocaba, SP

2019

Duarte, Juvenal José

Influência dos textos de notícias na queda de preços do mercado de ações brasileiro / Juvenal José Duarte. -- 2019.
114 f. : 30 cm.

Dissertação (mestrado)-Universidade Federal de São Carlos, campus Sorocaba, Sorocaba
Orientador: Sahudy Montenegro González
Banca examinadora: Sahudy Montenegro González, Deborah Silva Alves Fernandes, Katti Faceli, César Cruz Júnior
Bibliografia

1. Mineração de textos. 2. Notícias em português. 3. Mercado de ações brasileiro. I. Orientador. II. Universidade Federal de São Carlos. III. Título.

Ficha catalográfica elaborada pelo Programa de Geração Automática da Secretaria Geral de Informática (SIn).
DADOS FORNECIDOS PELO(A) AUTOR(A)
Bibliotecário(a) Responsável: Maria Aparecida de Lourdes Mariano – CRB/8 6979

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências em Gestão e Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Juvenal José Duarte, realizada em 14/08/2019:

Profa. Dra. Sahudy Montenegro González
UFSCar

Profa. Dra. Deborah Silva Alves Fernandes
UFG

Profa. Dra. Katti Faceli
UFSCar

Prof. Dr. José César Cruz Júnior
UFSCar

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Deborah Silva Alves Fernandes e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Profa. Dra. Sahudy Montenegro González

Em memória do meu pai, José de Jesus Rodrigues Duarte, a quem dedico este trabalho.

Agradecimentos

Agradeço

a Deus, por me dar sabedoria para enfrentar os desafios;
a minha família por me dar resiliência, incentivo e apoio nas dificuldades;
a minha orientadora, Sahudy Montenegro González, pela paciência e inúmeros ensinamentos compartilhados;
aos amigos Thiago Macedo e Maurizio Ruzzi, pelas conversas e sugestões sobre o tema;
e aos professores Tiago Agostinho de Almeida e José César Cruz Júnior, pelos conselhos.

"É necessário ter um conhecimento considerável para se dar conta de quão grande é a sua própria ignorância". (Thomas Sowell)

Resumo

Antecipar perdas financeiras e agir para evitá-las, ou ao menos amenizá-las, é um dos grandes desafios de qualquer investidor. Pela disponibilidade de dados e facilidade de aplicação, a análise técnica se difundiu rapidamente. Contudo, a evolução tanto no poder computacional como nas técnicas de mineração de texto, possibilitam hoje a análise de indícios comportamentais dos investidores em dados ainda pouco explorados: textos. Este trabalho apresenta um estudo sobre o mercado de ações brasileiro e sua relação com os meios de comunicação nacionais, com o enfoque na busca automática de padrões interpretáveis relacionados à queda de preços, através de algoritmos de aprendizagem de máquina. Foram executados seis experimentos para analisar a possibilidade de prever quedas automaticamente. Em sequência, foram estudados casos particulares em busca de explanações dos classificadores que justifiquem as predições. Os resultados apontam que as técnicas de mineração de texto se sobressaem às estratégias tradicionais na previsão de quedas, contudo a obtenção de explanações é limitada em função da complexidade dos classificadores e da alta dimensionalidade do vocabulário.

Palavras-chaves: Mineração de textos. Notícias em português. Mercado de ações brasileiro. Previsão de preços.

Abstract

Forecasting financial losses and making decisions to avoid or reduce them has been a challenge for every investor. On one hand, due to the availability of data and its simple implementation, technical analysis methods have been quickly gaining supporters. On the other hand, modern computers processing power together with advances in text mining provides the opportunity to explore the investor's behaviors in new data types: textual. This research evaluates the relationship between the Brazilian stock market and news published on national midia, focusing on automatic search for patterns related to down movements using machine learning algorithms. Six experiments were performed to analyze the possibility of predicting price falls automatically, followed by case studies in the search of explanations from the classifiers that justify the predictions.The results show that text mining based approaches overcome traditional strategies when forecasting losses, but the underlying patterns understanding is limited due to the complexity of the classifiers and high dimensional vocabulary.

Key-words: Text Mining. Brazilian Portuguese News. Brazillian Stock Market. Price forecasting.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AD	Árvores de Decisão
FA	Floresta Aleatória
RL	Regressão Logística
RNA	Redes Neurais Artificiais
MLP	*Multilayer Perceptron*
KNN	*K-Nearest Neighbors*
SVM	*Support Vector Machines*
SVM-L	*Support Vector Machines* com kernel Linear
NB-G	Naive Bayes Gaussiano
NB-M	Naive Bayes Multinomial
TF	*Term Frequency*
TF-IDF	*Term Frequency - Inverse Document Frequency*
BoW	*Bag of Words*
NLP	*Natural Language Processing*
PLN	Processamento de Linguagem Natural
HEM	Hipótese de Eficiência de Mercado
B&H	*Buy & Hold*
EMA	*Exponencial Moving Average*
MME	Média Móvel Exponencial
F1	F-Medida

Sumário

Prefácio

Problema recorrente, crises financeiras datam desde o sobrepreço de tulipas na Holanda em 1637, passando pela quebra da bolsa de Wall Street em 1929 até a recente bolha imobiliária norte americana, em 2008 (2). Para a crise de 2008, por exemplo, estima-se que a recessão de 18 meses tenha impulsionado a queda do produto interno bruto (PIB) americano em 4,3% e dobrado a taxa de desemprego de 5% para 10%, apresentando-se como a pior recessão desde a Segunda Guerra Mundial (3).

A possibilidade de antecipar movimentos futuros nos preços, dado o conhecimento disponível, é uma questão largamente estudada, mas ainda sem resposta conclusiva. Estudos empíricos apontam tanto sinais de aleatoriedade quanto padrões recorrentes nas variações (4). A hipótese mais aceita é a de que a possibilidade de prever os movimentos nos preços varia de acordo com o mercado observado, podendo este ser mais ou menos eficiente (5).

Se os investidores possuem acesso a toda informação disponível sobre as empresas e as interpretam racionalmente de maneira instantânea, este mercado apresenta *eficiência informacional* (5). Nesse caso, a previsão de preços é muito difícil, pois tão logo uma tendência é identificada os preços são imediatamente ajustados, anulando oportunidades de arbitragem (ganhos sem risco).

A observação de preços históricos e sua análise costuma ser a base da grande maioria de técnicas de predição. Abordagens quantitativas costumam ser empregadas, contudo a Economia é classificada como uma ciência humana, pois preços estão diretamente ligados aos sentimentos e reações das pessoas envolvidas nas negociações (6).

O ramo da Economia Comportamental estuda como fatores humanos se relacionam com as decisões tomadas em negociações, em especial o fator emocional (7, 6). Contraposta à hipótese de eficiência de mercado (5), as seguintes questões de pesquisa podem ser levantadas:

1. Para o mercado de ações brasileiro, é possível evitar perdas com estratégias automáticas baseadas em mineração de texto de notícias em português?
2. É possível obter dos classificadores explanações de modelo ou instâncias que descrevam os eventos por trás das depreciações?

Apesar de ser difícil responder de maneira conclusiva, respostas parciais para estas questões podem ser obtidas da observação de padrões históricos. Existem anomalias de mercado já conhecidas e racionalizadas (4), contudo encontrar novos padrões manualmente demanda longas análises em inúmeros conjuntos de dados. Dada esta limitação, a mineração

de texto se apresenta como solução, pois abordagens manuais não conseguem processar todo o volume de publicações tão rápido quanto seria necessário, seja para obter oportunidades ou evitar perdas no mercado financeiro (8).

Uma das fontes de informação mais usadas ainda hoje são notícias. Eventos históricos e estudos empíricos suportam a hipótese de que, se não determinantes, as notícias veiculadas exercem grande influência sobre a decisão de investidores, sobretudo quando são negativas (6).

A tentativa de formular um modelo capaz de prever preços futuros a partir de informação textual vem sendo explorada com relativo sucesso através da aplicação de algoritmos clássicos de classificação (9, 8, 10, 7). Entretanto, alguns pontos de melhoria são observados:

- A maioria dos estudos exploram mercados internacionais, com uso de textos em inglês (9, 10, 7). Pesquisas sobre o mercado nacional e conteúdo em português são raras (11, 12, 13).
- Os trabalhos em geral focam em modelos para otimizar acertos na predição da direção do preço, porém não são exploradas possíveis explanações para fundamentar as predições.
- Naïve Bayes, Máquina de Vetores de Suporte e Redes Neurais Artificiais são os algoritmos mais comuns, outras técnicas como K Vizinhos Mais Próximos e Árvores de Decisão são pouco estudadas, mesmo estas sendo potencialmente mais interpretáveis.
- Modelos são geralmente avaliados quanto à taxa de acerto das predições, mas não em termos financeiros. O retorno e a perda financeira são métricas importantes para avaliar a viabilidade dos algoritmos como estratégias, mas ainda assim costumam ser negligenciadas.
- As pesquisas normalmente se limitam à avaliação de poucas ações, de segmentos de negócio limitados.
- Apesar de apresentar melhores resultados em relação a antecipação de valorizações, nenhuma pesquisa dá ênfase à previsão de perdas.

O desafio de prever quedas de preço de ações, sobretudo fortes ondas de desvalorização, é interessante a entidades desde governos e investidores até empresas e fundos de pensão. Além da previsão, entender os eventos associados a essas variações fornece indícios do comportamento dos investidores, provendo melhor compreensão do contexto e embasamento às predições efetuadas.

Objetivos

Esta tese levanta duas questões principais de pesquisa, são elas:

1. Para o mercado de ações brasileiro, é possível evitar perdas com estratégias automáticas baseadas em mineração de texto de notícias em português?
2. É possível obter dos classificadores explanações de modelo ou instâncias que descrevam os eventos por trás das depreciações?

O objetivo da dissertação é tratar as perguntas levantadas, avaliando a relação entre as notícias públicas, veiculadas nos principais meios de comunicação nacionais, e oscilações negativas no preço de ações. Como propósito secundário, buscam-se padrões que descrevam os eventos responsáveis pelas depreciações.

Metodologia

Em um primeiro momento busca-se estudar se variações de preço no mercado brasileiro podem ser antecipadas por algoritmos de aprendizagem de máquina, sob enfoque de problema de classificação e treinados com texto de notícias em português, com desempenho semelhante aos estudos internacionais.

As notícias públicas são confrontadas com o preço negociado das 64 ações que compõem o índice Ibovespa, individualmente, para variados horizontes de tempo, a fim de buscar classificadores capazes de prever quedas de preço automaticamente. Métricas financeiras foram aplicadas na avaliação dos classificadores, além das quantitativas convencionais: Revocação, Precisão e F-Medida.

Além dos principais classificadores abordados na maioria das pesquisas (Naïve Bayes, *SVM* e RNAs), são avaliados comparativamente Árvore de Decisão, Floresta Aleatória, Regressão Logística e K Vizinhos Mais Próximos. Os classificadores são ainda contrapostos às estratégias tradicionais de operação *Buy & Hold* e Médias Móveis Exponênciais como *baseline*.

Em seguida, os modelos obtidos são analisados em relação à informação que estes proveem para suportar as predições fornecidas. O foco principal é identificar palavras ou expressões que possam agir como gatilho para desencadear um grande volume de vendas de ações. É analisada também a possibilidade de encontrar semelhanças entre datas recentes em relação a períodos históricos que antecederam quedas de preço, através da averiguação de proximidade proporcionada pelo algoritmo *KNN*.

Estrutura da dissertação

Como um trabalho multidiciplinar, a dissertação inicia apresentando os conceitos técnicos em duas seções principais: Mercado de Ações (no Capítulo 1) e Classificação de Textos (no Capítulo 2). Enquanto a primeira aborda o problema de um ponto de vista de negócio, a segunda discorre sobre as técnicas computacionais envolvidas na análise preditiva.

O Capítulo 1 começa discutindo o funcionamento do mercado de ações, seus agentes e sua organização. O cálculo de retornos é abordado em sequência, na Seção 1.1, descrevendo suas principais metodologias e propriedades. Na Seção 1.2 são apresentadas as discussões, sob o ponto de vista econômico, acerca da viabilidade do uso de técnicas preditivas. Por fim, são abordadas em 1.3 duas estratégias tradicionais de operação para compra e venda de papeis, sendo a primeira com embasamento fundamentalista (na Seção 1.3.1) e a segunda oriunda da análise técnica (na Seção 1.3.2).

O Capítulo 2 introduz as discussões computacionais apresentando os principais conceitos envolvidos na mineração de texto, bem como uma organização de etapas do processo. As fases de preparação dos dados são introduzidas pela avaliação da representação em 2.2, seguida das principais normalizações aplicáveis a dados textuais em 2.1 e, por último, apresentando estratégias de redução de dimensionalidade em 2.3, com métricas aplicáveis à seleção por filtro. Em 2.4 são abordadas metodologias consagradas para obtenção de classificadores, debatendo o funcionamento e as principais características de cada uma. O tema se encerra apresentando fundamentos científicos em torno da interpretação de padrões, obtidos através de aprendizagem de máquina, na Seção 2.5.

O Capítulo 3 são discutidos estudos semelhantes e suas abordagens para resolução do problema. Em 3.1 apresenta-se um sumário comparativo entre as principais questões de pesquisa e como estas são conduzidas nos trabalhos levantados. O capítulo se encerra com a contextualização da proposta em 3.2, reiterando as hipóteses, objetivos, diferenciais, principais desafios e limitações deste estudo.

O Capítulo 4 introduz as etapas da metodologia adotada nos experimentos e o protocolo de testes. É discutido o escopo do problema e as janelas temporais de interesse, os dados desde a captura, transformação, integração e representação computacional, a organização do treinamento e teste de modelos e as medidas de avaliação usadas na verificação de desempenho.

Os experimentos são exibidos no Capítulo 5, sendo que 5.1 reúne a sequência de experimentos responsáveis por otimizar as predições dos modelos, e 5.2 busca obter dos classificadores explanações de modelo e instância que justifiquem as predições efetuadas.

O trabalho se encerra com a revisão das questões de pesquisa e as conclusões finais baseadas na análise experimental. Estudos futuros são propostos amparados nos temas

mais promissores observados ao longo da pesquisa.

1 Mercado de ações

Como uma maneira de arrecadar mais investimentos para seus negócios, muitas empresas optam pela abertura de capital, tal que suas cotas se tornam um ativo financeiro negociável (14).

O detentor de ações de uma empresa, no momento da compra do ativo, assume a posição de proprietário de uma cota da companhia, ora com acesso aos lucros através de dividendos, ora apenas com a garantia de posse de um bem passível de valorização ou depreciação (14).

As negociações de papéis são organizadas através de um mercado, o qual aproxima vendedores de compradores e regula as transações entre estes. De maneira mais genérica, são definidos como mercados financeiros os comércios de produtos financeiros diversos, desde contratos de renda fixa, derivativos a contratos de balcão, tendo como ativos desde *commodities* até futuros.

O mercado de ações é uma ramificação do mercado financeiro, tendo como ativo objeto apenas papéis de companhias de capital aberto. As transações de um mercado de ações podem ocorrer por meio de bolsas de valores ou mercados de balcão, sendo o primeiro destinado a contratos padronizados, também chamados *plain vanilla*, e o segundo para contratos de termos customizados (14).

Enquanto as bolsas de valores são responsáveis por organizar as ordens de compra e venda conduzindo ao fechamento de negócios, corretoras são responsáveis por atuar como intermediárias entre os investidores e as bolsas de valores, fornecendo uma interface para estes emitirem suas ordens de compra e venda. As operações normalmente envolvem custos, desde taxas de operação e administração até a incidência de impostos (14). De maneira simplificada, a iteração entre os agentes envolvidos no mercado de ações é exibida na Figura 1.

A definição de preços do mercado de ações é gerida através da lei de oferta e demanda. Ao submeter uma ordem de venda por um preço determinado, o investidor só tem seu negócio concretizado se existir uma ordem de compra com preço equivalente, e vice e versa (15). Corretoras oferecem também a opção de emitir ordens com características mais flexíveis, como ordens realizáveis com preço até certo limite (14).

As negociações ocorrem em dias comerciais, em horário comercial (14). A cotação de abertura dos ativos define o preço a ser negociado no início do pregão, ao passo que o preço de fechamento representa o valor do ativo na última transação antes do fechamento da bolsa.

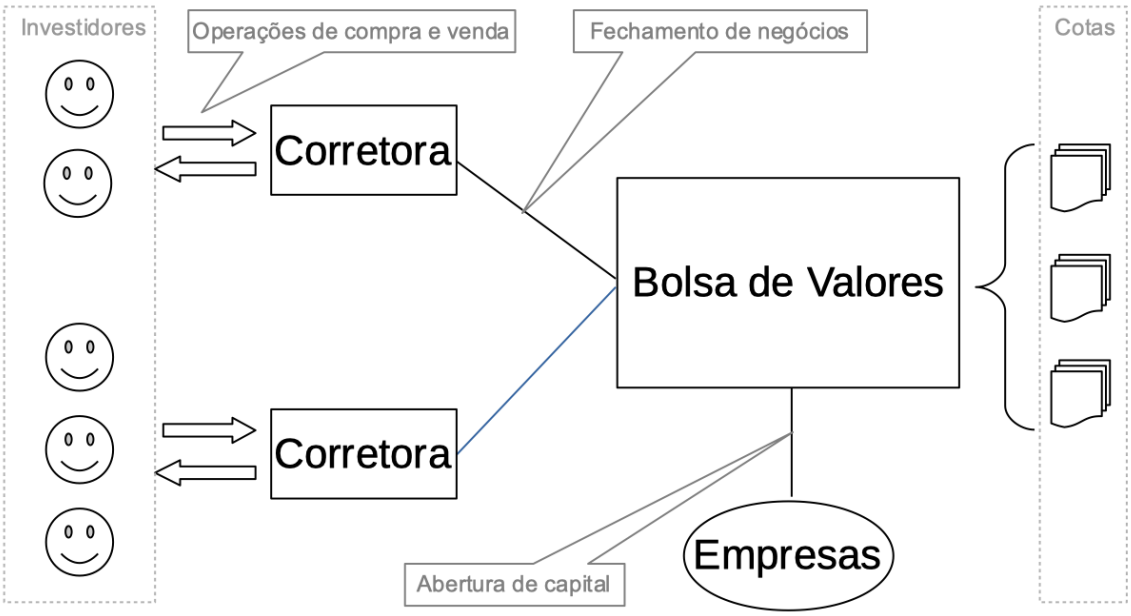


Figura 1 – Simplificação dos agentes no mercado de ações e suas iterações.

Fonte: Produzido pelo autor

Em terminologia financeira, os investidores detentores de papéis têm sua posição chamada de ”comprada” na ação, ou posição ”longa” (do inglês *long*), operadores que não possuem o papel em carteira tem sua posição chamada ”vendida”, ou ”curta” (do inglês *short*).

Ao longo do dia as informações de preços negociados, bem como o volume de negócios fechados, são armazenadas com o intuito de fornecer informações aos participantes do mercado.

1.1 Cálculo de retornos

Investimentos, de modo geral, são avaliados de maneira comparativa a fim de determinar quais produtos oferecem os melhores benefícios associados ao menor risco. Os benefícios são comumente examinados em termos de ganho de capital, descrevendo o lucro obtido através do investimento de um montante inicial após o encerramento do período de vigência do contrato (16).

A taxa de retorno de um investimento descreve seu desempenho de maneira mais genérica, podendo esta ser medida de modo absoluto (em unidade de reais, por exemplo) ou percentual. Um investimento com taxa de retorno positiva é um investimento lucrativo, enquanto um investimento resultante em perdas tem taxa de retorno negativa. Dependendo da liquidez e do tipo de contrato negociado, a frequência de avaliação do retorno financeiro pode ser menor ou maior.

No mercado de ações a liquidez dos papéis está associada ao volume de negócios: desde que haja ao menos duas partes interessadas em efetuar a transação sob os mesmos termos, haverá negócio. Assim, os retornos podem ser calculados de maneira intradiários, baseados nos preços praticados durante o dia, no intervalo diário, usando a diferença entre o preço de abertura e de fechamento, ou períodos mais longos.

O retorno multiplicativo é uma das metodologias mais comuns. Dado um intervalo de tempo Δ_t e desconsiderando possíveis dividendos, a taxa retorno de uma ação no período é dado pelo quociente entre o preço final e o preço inicial, decrescido de 1, conforme a equação 1.1 (17).

$$Retorno(\Delta_t) = \left(\frac{Preço_{final}}{Preço_{inicial}} \right) - 1 \tag{1.1}$$

A taxa de retorno é útil por expressar diretamente a magnitude de ganhos ou perdas, com limiar entre eles fixado em 0. Em termos matemáticos, contudo, taxas de retorno costumam ser representadas por fatores, com centro em 1, descrevendo a fração do capital inicialmente investido ao final do período. Com o fator de retorno representado pela função $F()$, a fórmula é dada em 1.2

$$F(\Delta_t) = Retorno(\Delta_t) + 1 \tag{1.2}$$

O fator de retorno final de um investimento pode ser calculado tanto diretamente, como o quociente entre o preço final e o inicial, quanto como uma composição de retornos individuais. Assumindo que o intervalo Δ_t pode ser subdividido em N sub intervalos passíveis de averiguação de fatores de retornos, o fator de retorno final do investimento será dado pela equação 1.3.

$$F(\Delta_t) = \prod_{i=1}^{n} F(\Delta_i) \tag{1.3}$$

O produtório de fatores na equação 1.3 é também chamado acúmulo de retornos ou *accrual*. O uso de *accrual* é especialmente interessante quando empregado na análise de investimentos de renda variável, onde a taxa de retorno não é constante, e também quando a posição de investimento pode se alterar ao longo do período analisado.

Uma outra maneira de observar o desempenho de ações é avaliá-las em termos de um índice base de mercado. No Brasil, o índice Ibovespa mede o desempenho de uma carteira composta das 64 ações com maior volume e liquidez de negociações. A grosso modo, o índice busca avaliar o desempenho médio do mercado, e é usado por economistas como um indicador de comportamento do mercado acionário como um todo [1].

[1] Detalhes da composição e metodologia acessados em: <http://www.bmfbovespa.com.br/pt_br/produtos/indices/indices-amplos/indice-bovespa-ibovespa.htm>

Ao avaliar o retorno de uma ação através das oscilações de preço somente, não só o papel é avaliado, mas também o mercado financeiro como um todo, a situação econômica. Uma ação pode ter desempenho positivo por eventos econômicos não necessariamente ligados aos negócios da empresa, o mesmo ocorre para depreciações. O risco sistemático se refere ao risco ou probabilidade de quebra de um sistema como um todo, ao contrário da quebra de partes e componentes individuais (18).

Retornos anormais visam isolar os fatores econômicos dos eventos puramente relacionados à operação da empresa. Para o caso de retornos ajustados ao mercado (*Market Adjusted Returns*), os retornos de um índice base de mercado, ou índice de referência, são subtraídos do retorno da ação, produzindo uma métrica de performance da ação em relação a média do mercado (17). A equação é dada em 1.4.

$$FatordeRetornoAnormal(\Delta_t) = \prod_{i=1}^{n}(FatordeRetornoAção(\Delta_i) - FatordeRetornoÍndice(\Delta_i)) \tag{1.4}$$

O uso de retornos anormais isola os movimentos de preço da ação das variações de mercado, permitindo uma análise mais focada de seus investidores. Por outro lado, por estar indexado a uma taxa flutuante, não necessariamente retrata lucros ou perdas, mas sim a performance da ação em relação ao índice.

1.2 Previsibilidade

O mercado de ações ganhou muitos adeptos ao longo dos anos, mas ainda hoje é visto como loteria por muitos de seus participantes, sobretudo pela volatilidade dos preços e imprevisibilidade de seus eventos.

Boa parte das teorias econômicas assumem que o preço dos ativos se comporta de maneira aleatória, ou em função de tantas variáveis que sua previsão é completamente inviável (7). A Teoria de Passeios Aleatórios define que as variações de preço de um papel se desenvolvem através de movimentos arbitrários, independentes e subsequentes, tal que sua previsão é impraticável (5). As variações estocásticas de preço descreveriam Movimentos Brownianos, com distribuição gausssiana e desvio padrão dado pela volatilidade do ativo.

Muitos estudos empíricos, por outro lado, apontam contra exemplos com variações previsíveis ao longa da história do mercado acionário. Os contra exemplos se manifestam sobretudo em padrões conhecidos, chamados anomalias de mercado, documentados e testados empiricamente. As anomalias variam desde padrões de calendário como o "Efeito Janeiro", fundamentalistas como o "Efeito Tamanho" até observações técnicas, como "Quebras de Faixa de Negociação" (4, 19).

Fama propôs em (5) a análise da previsibilidade associada à eficiência informacional dos mercados, tal que mercados mais eficientes se comportariam de maneira mais aleatória, e, quanto maior a ineficiência, maiores as chances de sucesso em predições e arbitragem. A Hipótese de Eficiência de Mercado (HEM) defende que, se toda informação relevante se reflete completamente nos preços, o mercado em questão pode ser considerado eficiente (5).

Assumindo que as informações são sempre absorvidas de maneira imediata, a Hipótese de Eficiência de Mercado (HEM) se descreve de três formas, baseado no tipo de informação:

- **Forma Fraca:** Os preços históricos são acessíveis e assimilados por todo o mercado.
- **Forma Semiforte:** Tanto os preços quanto toda informação pública é acessível e absorvida pelo mercado.
- **Forma Forte:** Informações históricas, públicas e também privadas são imediatamente processadas e refletidas nos preços.

A HEM alega que predições com oportunidades de ganho, ou arbitragem, não podem ser obtidas através de informações descritas em suas formas. A argumentação se baseia no fato de que, se as informações forem imediatamente absorvidas, os preços também são imediatamente ajustados, sem a possibilidade de arbitragem. Desta forma os preços são constantemente revisados a um valor neutro, caracterizando uma concorrência leal (do inglês *fair play*).

Desde sua proposta inicial, a Hipótese de Eficiência de Mercados se tornou a base para estudos em previsibilidade. Testes empíricos evidenciam tanto a eficiência quanto o contrário, dependendo do mercado em questão e, sobretudo, do período avaliado. Análises dos preços mostram que um mesmo mercado pode se comportar de maneira eficiente em um intervalo, mas de maneira irracional em outros ciclos.

Santos avalia em (19) que não só o mercado reage de maneiras distintas à informação, mas também diferentes grupos de agentes tendem a apresentar assimetria informacional. Grupos de investidores teriam maior capacidade de antecipar movimentos nas cotações, aparentando assimilar melhor as informações disponíveis, enquanto outros grupos agiriam de maneira menos racional, ou menos esclarecida. A existência de alguns agentes racionais no mercado, contudo, não se mostra suficiente para garantir que o equilíbrio se estabeleça sobre os preços (20, 21).

Estudos recentes têm explorado o fator psicológico como uma das principais forças que regem as negociações. Apesar do rigor formal e matemático, a economia ainda é o estudo de relações humanas, uma ciência humana, logo os sentimentos e percepções dos agentes que compõe o mercado financeiro têm efeito direto sobre suas decisões (22).

O ramo de Economia Comportamental investiga fatores psicológicos, cognitivos, emocionais, culturais e sociais envolvidos no processo de decisão de operadores financeiros (23). As teorias investigam eventos históricos evidenciando reações irracionais nas cotações e suas motivações, alguns casos são:

- **Bolhas:** Shiller conduziu em (22) um extenso estudo sobre casos de bolhas ocorridas na bolsa de valores norte americana, apontando comportamentos irracionais e, portanto, justificando sua ineficiência.
- **Reações exageradas:** Em (24) e (20) são discutidos casos de reações desproporcionais apresentadas nos preços, dados eventos ocorridos. Investidores não agiriam de maneira lógica ao interpretar informações, mas sim de maneira emocional.
- **Anomalias:** Em (4) são discutidas anomalias de mercado evidenciadas em períodos históricos no mercado acionário brasileiro. Padrões recorrentes motivados por razões conhecidas indicam condições em que preços podem ser antecipados com certo sucesso.
- **Aversão a riscos:** Perdas e ganhos financeiros são administrados de maneira diferentes pelos investidores, sendo que prejuízos costumam ter maior impacto emocional (25). Apesar da teoria clássica assumir que o valor esperado para o preço de uma ação em t_{+1} é o próprio valor em t, por martingale, as evidências empíricas demonstram que muitas vezes as variações de preço não são normalmente distribuídas, tão pouco descrevem um processo de martingale, sobretudo por que as observações não se mostram independentes.
- **Autoconfiança excessiva:** Em (26) são discutidas finanças comportamentais em relação ao mercado brasileiro. Dentre os tópicos cobertos, o autor cita que, entre os investidores, a maioria considera acima da média sua habilidade para vencer o mercado.

Estudos como (27) e (20) discutem de maneira antagônica as estratégias mais eficazes e motivações comportamentais que modelariam a ocorrência de retornos anormais, contudo ambos corroboram ao reconhecer que os preços estão sujeitos a ineficiência em razão de fatores humanos.

Se o mercado se comporta de maneira não eficiente ou irracional, mesmo que por vezes e períodos curtos, isso implica que agentes com maior capacidade de absorver, interpretar e tomar decisões baseado nas informações teriam maior chances de sucesso. O período entre os eventos e o reflexo no preço proporciaria oportunidades de arbitragem, especialmente a aqueles com maior capacidade de acesso e interpretação de informações (27).

O primeiro desafio ao tentar antecipar cotações é que o mercado de ações é adaptativo. Oportunidades de ganho tendem a ser anuladas assim que estas se tornam conhecidas. A segunda dificuldade se encontra no escopo abrangente de análise, com ínumeros mercados, ativos, janelas de tempo e técnicas diferentes. Padrões raramente são generalizáveis, tão pouco permanentes, demandando revisão empírica constante.

Em termos de predição, as abordagens se classificam em duas famílias: Análise Técnica e Análise Fundamentalista. Na análise técnica supõe-se que, conforme a teoria de Fama (5), toda informação disponível se reflete no preço negociado, portanto a observação de padrões históricos deve fornecer indícios dos comportamentos futuros do papel. A Análise Fundamentalista, por outro lado, se empenha em avaliar o valor intrínsico da ação baseado em indicadores tanto qualitativos como quantitativos da empresa, tais como lucro projetado, valor e composição dos ativos e passivos, fatia de mercado e possibilidade de expansão (28).

Abordagens técnicas costumam obter mais adeptos por facilitar a automação de análises, graças a oferta de dados estruturados e algoritmos com amplas opções de implementação. Para técnicas fundamentalistas, as análises costumam ser mais subjetivas, demandando especialistas na tarefa, além de dados não uniformes e métodos na maioria das vezes manuais.

Uma vez estabelecida sua percepção do mercado, da ação e de seu prospécto futuro, o investor tenta estabelecer uma estratégia de operação. As estratégias são responsáveis por sumarizar as informações em uma decisão, mais especificamente, em uma ordem de compra ou venda.

1.3 Estratégias de operação

Como discutido no capítulo anterior, as decisões do investidor estão muito ligadas à sua percepção de riscos e oportunidades. Através de suas análises, embasadas por seu conhecimento, este toma a decisão de quanto de seu dinheiro investir e em quais empresas.

A escolha das empresas e do capital a ser investido em cada uma delas é chamada otimização ou seleção de portifólio (29). Por outro lado, decidir sobre o momento oportuno para compra e a venda dos papeis normalmente requer um planejamento, a estratégia de operação.

Apesar da grande variedade de estratégias disponíveis, desde fundamentalistas a técnicas, passando por automáticas e manuais, alguns métodos são mais difundidos entre os participantes do mercado acionário. A seguir são introduzidas duas técnicas simples, mas largamente utilizadas.

1.3.1 *Buy & Hold*

Buy & Hold costuma ser considerada um *baseline* para comparação de estratégias. Sua lógica envolve comprar um ativo e mantê-lo em carteira por um tempo determinado, sem negociá-lo. O investidor que adota esta técnica tem perspectivas positivas de longo prazo com relação aos papéis que negocia.

A técnica é considerada uma estratégia passiva, com poucas operações e pouco sensitiva a oscilações de curto prazo. As análises se focam muito mais nas tendências do que em eventos isolados.

Como as operações são pontuais, o retorno financeiro desta técnica pode ser avaliado como um único período Δ_t, sendo dado pelo quociente entre o preço na data de venda e o preço da compra.

1.3.2 Médias Móveis Exponenciais

O uso de médias móveis tenta detectar anomalias na tendência de preços negociados. Para isso, é determinada uma janela de observação de n dias anteriores à data alvo e então calcula-se a média dos preços. Para o caso específico de médias móveis exponenciais, é atribuído mais importância ao preço mais recente 1.5.

$$m[t] = (m[t] - m[t-1]) \times (\frac{2}{n+1}) + m[t-1] \tag{1.5}$$

Na fórmula 1.5, $m[t]$ representa a t-ésima média móvel, composta pela média anterior $m[t-1]$ acrescida da diferença entre esta e o preço mais recente, descontado pelo fator $2/n+1$, onde n é o tamanho da janela de dias.

O efeito obtido ao calcular a série de médias móveis é suavizar a série de preços, de forma que variações atípicas se destacam em relação a série suavizada. Como estratégia de operação, avalia-se o preço negociado em relação a média dos dias anteriores e, caso este seja superior, a venda deve ser efetuada, do contrário deve-se assumir a posição de compra.

Em resumo, a estratégia investiga, baseada na tendência de preços anteriores, se o valor negociado atual é barato ou caro.

1.4 Considerações finais

No capítulo 1 foram discutidas as dinâmicas do mercado de ações e apresentada a teoria por trás da formulação dos preços: a lei da oferta e demanda. Apesar dos preços se comportarem de maneira aparentemente complexa e aleatória, a discussão apresentada em 1.2 apresenta vários pontos que corroboram com a hipótese de que há oportunidades para investidores com maior capacidade de capturar e processar informações, desde que o

mercado apresente baixa eficiência informacional e as ordens sejam submetidas em tempo hábil.

Através da avaliação de novas estratégias por meio de seus retornos financeiros, como apresentado em 1.1, é possível avaliar os riscos e lucros envolvidos na operação. Aplicada aos classificadores obtidos através de técnicas de aprendizado de máquina, estas métricas permitem que estes modelos possam ser comparados a estratégias tradicionais, como as abordagens técnicas e fundamentalistas, quanto a sua viabilidade.

No capítulo seguinte discutimos os conceitos necessários para formulação de estratégias modeladas como um problema de classificação de textos, com algoritmos clássicos como proposta de solução.

2 Classificação de textos

De acordo com (30), a ideia de mineração de textos de maneira automática com o auxílio computacional teve início em 1958, com o trabalho sobre a sumarização automática de textos literários proposto em (31). Por volta do ano 2000, já com o acesso à web mais difundido, a área de pesquisa ganhou maior notoriedade e aplicações, incentivada pelo volume de dados e poder computacional disponíveis.

As tarefas de mineração de texto são melhor organizadas segundo seus objetivos, tendo como categorias principais a classificação, agrupamento, mineração de regras de associação, sumarização e detecção de assunto (10). Os algoritmos aplicados nestas tarefas possuem origem normalmente na área de mineração de dados estruturados, empregados a documentos após uma etapa de conversão do conteúdo em texto para uma representação estruturada (32).

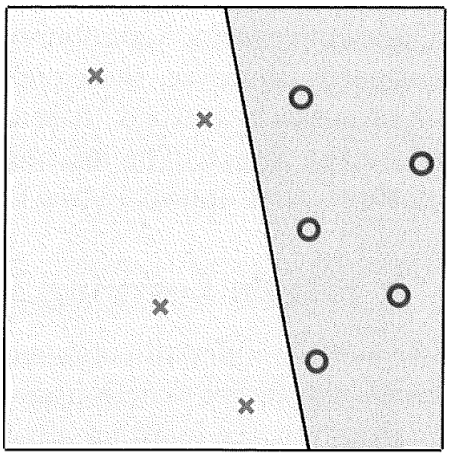
Quanto ao modelo de aprendizagem, as duas formas principais são supervisionada e não-supervisionada. No primeiro caso, que engloba por exemplo os problemas de classificação, os modelos são treinados através de exemplos e depois aplicados de maneira preditiva na análise de novos registros. No segundo caso, usado por exemplo em análises de agrupamento, os algoritmos são aplicados sobre os dados em busca de padrões descritivos. Uma terceira maneira, frequentemente empregada em robótica, é o aprendizado por reforço, onde o algoritmo inicia sem conhecimento prévio e constrói seu modelo baseado em estímulos e penalizações atribuídos a suas decisões (33, 34).

O aprendizado supervisionado baseia-se na hipótese de aprendizagem por indução, que argumenta que uma função que aproxima os resultados de uma função ideal em um *dataset* grande o suficiente, também aproximará bem seus resultados em exemplos ainda não observados (35).

Classificação, categorização, ou ainda em caso específico análise de sentimento, consiste na busca de um modelo capaz de determinar o valor de um atributo alvo baseado nos valores dos atributos de treino. Almeja-se, através do treino com exemplos já rotulados, obter fronteiras entre as classes tal que novos registros, ainda não classificados, possam ser rotulados por generalização (como nas figuras 2 e 3).

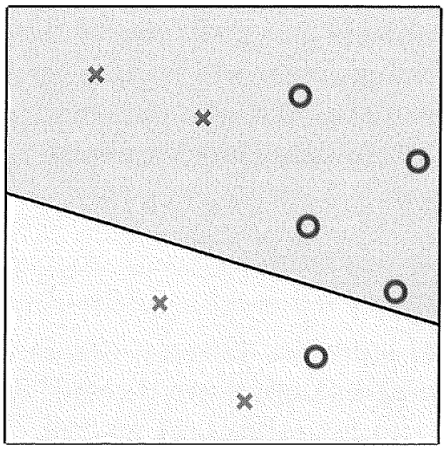
Um dos maiores problemas na busca de classificadores é o *trade-off* entre generalidade (*bias*) e representatividade (variância). Modelos muito genéricos (sub-ajustamento) não capturam padrões suficientes nos dados para atingir um bom desempenho em aplicações reais. Classificadores complexos (super-ajustamento), por outro lado, acabam por mapear tão bem os dados de treino que não são capazes de se adaptar a novos dados desconhecidos (33).

Figura 2 – Classificação bem sucedida.



Fonte: Traduzido de (34).

Figura 3 – Classificação mal sucedida.



Fonte: Traduzido de (34).

A avaliação dos classificadores se dá, na maioria dos casos, por medidas de acerto, sendo as mais comuns acurácia, revocação, precisão e F-medida. Estas medidas quantificam o desempenho do modelo computacional em relação ao sistema real, comparando as predições realizadas com os resultados reais.

Apesar de fundamental, a busca por padrões representa apenas uma das etapas finais de um processo maior chamado extração de conhecimento (36, 37). As etapas envolvidas na extração de conhecimento podem variar de acordo com o problema, a granularidade de detalhes e a abordagem, mas em geral envolvem os passos descritos na Figura 4.

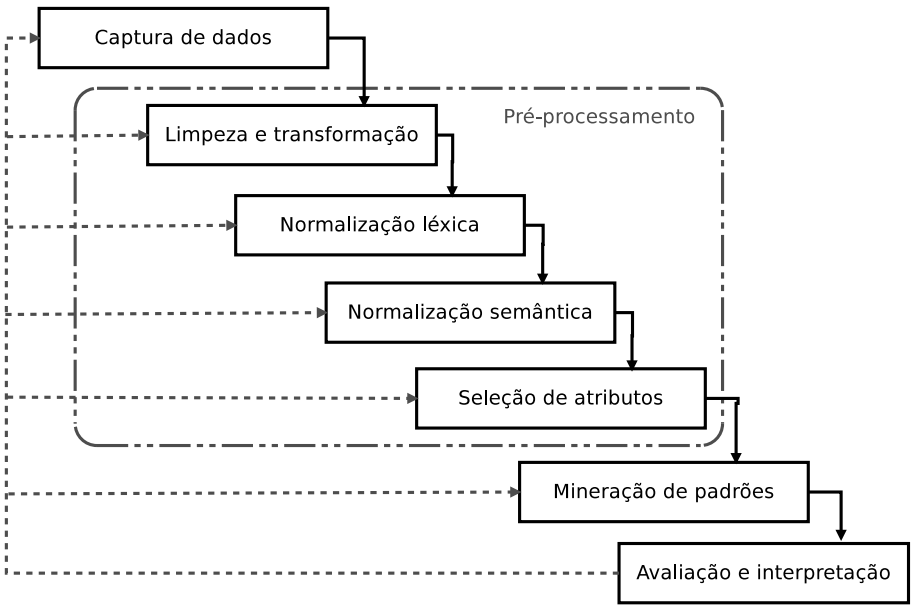


Figura 4 – *Pipeline* de tarefas de Mineração de Dados adaptadas para Mineração de Texto.

Fonte: Adaptado de (37).

A captura, ou seleção de dados, abrange a investigação das fontes de informação e quais dados podem ser recuperados, e de que maneira. Este processo é essencial pois define a viabilidade de se obter as respostas almejadas com base nos dados disponíveis.

As etapas que seguem podem ser resumidas como pré-processamento, contemplando a limpeza de ruídos nos dados originais, transformação para uma representação computacional adequada aos métodos de análise, aplicação de técnicas de normalização, tanto léxicas quanto semânticas, e por fim a seleção de atributos.

A mineração de padrões é a aplicação de técnicas de análise aos dados, as quais devem produzir informações relevantes para apoiar o processo de tomada de decisão. Com os padrões já encontrados, segue-se à etapa de avaliação de qualidade e intepretação, produzindo por fim o conhecimento a partir dos dados.

2.1 Normalização

Técnicas de normalização buscam otimizar os dados textuais de forma a obter melhores resultados nas análises de padrões. Enquanto as técnicas léxicas agem no filtro, correção de erros e padronização das palavras, as técnicas semânticas buscam enriquecer o sentido destas, agindo na desambiguação, tratamento de sinônimos e verificação de contexto.

Entre as técnicas mais comuns, encontra-se a padronização em caixa baixa. Seu emprego contribui para que palavras iguais não sejam tratadas como distintas em função de diferenças léxicas (1), como no exemplo em 2.1.

$$\{Escola, escola, ESCOLA\} \Rightarrow \{escola\} \tag{2.1}$$

Outra padronização com o mesmo propósito é a remoção de caracteres especiais, como em 2.2, a qual contribuí tanto para contornar adequações ortográficas quanto para o tratamento de erros de digitação (*typo*).

$$\{bilíngue, bilíngüe, bilingue\} \Rightarrow \{bilingue\} \tag{2.2}$$

Dentro da gramática de linguagens naturais, muitos termos como artigos, conjunções e preposições são usados mais pela organização do texto do que por seu valor semântico, agregando pouca ou nenhuma informação (1). Estas palavras costumam ser chamadas de *stopwords*, exemplos no idioma português são dados em 2.3.

$$\{em, de, na, a, se\} \tag{2.3}$$

Em (38) os autores defendem que *stopwords* ocorrem com tanta frequência nos documentos que seu poder de discriminar entre as classes alvo é irrisório. A remoção destes termos contribui tanto para a simplificação como para a redução de ruídos.

Muitas linguagens, como no caso do português, apresentam diversos tipos de flexão nas palavras, seja por tempo, voz, gênero, número etc. A flexão de termos normalmente envolve a adição de prefixos ou sufixos, adaptando o radical ao contexto.

Existem duas abordagens mais comuns para tratar a flexão de palavras, tal que diferentes grafias sejam agrupadas por seu significado comum. A primeira diz respeito a extração de radical (*stemming*), quando os sufixos são removidos até que o radical da palavra seja encontrado. O segundo caso é o uso de lematização, que busca traduzir o termo à sua forma canônica. Um comparativo entre as técnicas é dado na Tabela 1.

Tabela 1 – Comparativo entre regularizações de flexão por *stemming* e lematização. Exemplos extraídos de 1.

Palavra	***Stemming***	**Lematização**
felicíssimo	felic	feliz
segurados	segur	segurar
anoitecendo	anoitec	anoitecer

A extração de radical é mais simples e tem menor custo computacional, no caso do português pode ser implementada pelo algoritmo RSLP (1), mas em muitos casos pode conduzir a descaracterização da palavra original. A lematização produz resultados melhores, entretanto é mais complexa e seu custo computacional é maior devido à análise semântica (1).

Muitas vezes, quando o problema envolve a análise de elementos gramaticais no texto para a identificação de nomes ou eventos, por exemplo, é interessante etiquetar cada um dos *tokens* de acordo com sua classe gramatical, processo chamado de *Part of Speech Tagging*, ou simplesmente *POS Tagging* (32).

Linguagens naturais oferecem diferentes maneiras de expressão, possibilitando diversos sentidos em uma mesma palavra através de figuras de linguagem. A redução de ambiguidade no significado de palavras é um desafio no processamento de linguagens naturais (*NLP*), pois métodos computacionais dependem da formalização de conceitos para poder processá-los.

Dicionários e ontologias costumam ser aplicados na indexação semântica de termos, sendo o projeto *WordNet* um dos mais longos e famosos (32). Entretanto, um dos grandes problemas é que, após determinar os significados possíveis de um termo, é necessário definir qual o conceito mais se adequa ao sentido no contexto original.

Uma abordagem para o enriquecimento semântico de textos é proposta na pesquisa (39), sugerindo um processo de três etapas: 1. Normalização de palavras; 2. Expansão de conceitos; 3. Desambiguação de sentido. O processo busca padronizar a representação de conceitos semânticos formalizando as informações para a busca de padrões.

No trabalho (40), considera-se a engenharia de atributos como um passo fundamental para o reconhecimento de padrões via algoritmos de aprendizagem, contudo sua execução se torna ainda mais complexa por se tratar da análise de domínios específicos de cada *dataset*.

Apesar dos diversos tratamentos apresentados, e muitos outros existentes, discute-se muito qual o grupo de técnicas mais adequadas a cada problema. Em geral, a resposta deve ser explorada através de experimentação, ajustando o *trade-off* entre simplicidade, informação preservada, viabilidade e custo benefício caso a caso.

2.2 Representação computacional

Quando se trabalha com dados não estruturados, algumas definições se fazem necessárias para melhor compreensão do problema. Os documentos representam um registro de dados, cujo conteúdo normalmente se apresenta em texto em linguagem natural. A junção de vários documentos como um corpo de dados é referenciada na literatura como uma coleção de documentos ou *corpus*, que por sua vez pode ser analisada tendo como atributos (do termo inglês *features*) cada caractere, palavra, termo ou conceito individual (41). O processo de transformação de texto em atributos é chamado de segmentação ou *tokenização*.

Como bem definido em (42), "computadores entendem muito pouco do significado da linguagem humana", portanto estabelecer uma representação formal para este tipo de dado é indispensável. Apesar de existirem representações alternativas como em árvores ou grafos, o uso de espaços vetoriais para descrever os dados é ainda hoje a abordagem mais adotada (7).

Na representação por espaço vetorial cada termo presente no *corpus* se torna uma coluna, sendo que os respectivos valores atribuídos a estas são normalmente a ocorrência, em formato binário, ou a frequência, representando quantas vezes a palavra apareceu no texto (36). O uso de ocorrencias é adotado em sua implementação mais comum, *Bag of Words* (*BoW*).

A avaliação de frequência de termos pode ser feita pela métrica *TF* (*Term Frequency*), contudo outra técnica bastante difundida para determinar pesos para as palavras é o *TF-IDF* (*Term Frequency - Inverse Document Frequency*), o qual busca tratar a relevância relativa das palavras pelo quociente entre a frequência da palavra no documento

e o inverso de sua frequência em todo o corpus (36, 1).

Sob o enfoque do ramo de Recuperação de Informações (do inglês *Information Retrieval*), é avaliado quão bem os modelos descrevem os dados originais, além de quanto da informação primitiva é preservada na nova representação.

Quando considerada a frequência ao invés da ocorrência binária, pode-se dizer que são preservadas mais informações, o volume de ocorrências. A técnica *N-Gram* busca manter a relação semântica entre as palavras tratando termos consecutivos como únicos, onde a cesta de *tokens* {"aumento","de","impostos"} do *BoW* seria representada como {"aumento de impostos"} em *N-Gram*, para N=3.

2.3 Seleção de atributos

O objetivo da seleção de atributos pode ser sintetizado em identificar características importantes, descartando as irrelevantes e redundantes (37). Distinguir dentre os atributos disponíveis os mais relevantes ao problema é interessante, pois reduz complexidade, viés e esforço na obtenção de padrões. Em resumo, idealmente deve-se buscar o conjunto de atributos independentes que apresentam maior correlação com o atributo alvo para se obter os melhores resultados (40).

Existem diferentes métodos para desempenhar a redução de dimensionalidade, contudo estes podem ser resumidos em três categorias principais: 1. Filtro; 2. *Wrappers*; 3. *Embedded* (43, 44). Para o primeiro caso, por filtro, busca-se através de uma métrica de importância ordenar os atributos para depois filtrá-los. Os métodos *Wrappers* executam iterativamente o processo de aprendizagem, colhendo informações de desempenho e otimizando o subconjunto de atributos. Métodos *Embedded* são embutidos no próprio processo de aprendizagem (43).

Tanto as técnicas baseadas em *Wrappers* quanto *Embedded* costumam ser mais custosas computacionalmente por se tratarem de métodos iterativos. Por esse motivo, em casos de problemas complexos com grande volume de dados e alta dimensionalidade, estas abordagens facilmente se tornam inviáveis (1).

As técnicas baseadas em filtro são mais diretas, descartando as *features* cujas métricas estão abaixo de um limite ou simplesmente selecionando o *ranking* dos N melhores atributos. As medidas usadas no filtro normalmente avaliam a dependência dos atributos de treino entre si ou em relação ao atributo alvo, as métricas mais comuns são apresentadas a seguir.

2.3.1 Variância

Dada uma variável X, a variância $Var(X)$ caracteriza a dispersão de seus valores em torno de sua média, conforme a fórmula 2.4 (35). Em outras palavras variáveis com alta variância assumem valores mais heterogêneos, enquanto uma variável com $Var(X) = 0$ assume valores constantes.

$$Var(X) = E[(X - \mu_X)^2] \tag{2.4}$$

Medir a variância dos dados em um determinado atributo produz uma métrica simples, mas muito útil. Atributos constantes, por exemplo, onde o mesmo valor é atribuído para todos os registros, não agregam informação nenhuma ao modelo de classificação, pois não mostram alteração em seus dados independente da classe alvo.

No caso de mineração de texto, a remoção de termos com baixa variância é especialmente útil para redução de ruídos, muitas vezes gerados na fonte dos dados ou durante a captura.

2.3.2 Chi2 (χ^2)

A estatística χ^2 mede a falta de independência entre duas variáveis previamente assumidadas como independentes (45), sendo que dois eventos A e B são definidos como independentes se $Probabilidade(A|B) = Probabilidade(A)$ e $Probabilidade(B|A) = Probabilidade(B)$ (46).

Se $C = \{c_1, c_2, ..., c_n\}$ é o conjunto de possíveis classes e $A = \{v_1, v_2, .., v_n\}$ os valores possíveis no domínio do atributo A, definem-se os valores esperados e observados como nas expressões 2.5 e 2.6.

$$valor_{observado}(v_i, c_i) = count(A = v_i, C = c_i) \tag{2.5}$$

$$valor_{esperado}(v_i, c_i) = \frac{count(A = v_i) \times count(C = c_i)}{count()} \tag{2.6}$$

A partir dos valores auferidos da contagem de frequência, avalia-se o valor estatístico χ^2 pela fórmula 2.7 (46).

$$\chi^2(v_i, c_i) = \frac{(valor_{observado}(v_i, c_i) - valor_{esperado}(v_i, c_i))^2}{valor_{esperado}(v_i, c_i)} \tag{2.7}$$

As fórmulas 2.5, 2.6 e 2.7 dizem respeito a avaliação de um valor no domínio A em relação a uma classe no conjunto C apenas. Dadas as cardinalidades $\#A$ e $\#C$,

representando o número de valores possíveis nos conjuntos, o resultado do teste de independência entre o atributo A e o atributo alvo C é dado pela somatória em 2.8 (46).

$$\chi^2(A, C) = \sum_{i=1}^{\#A} \sum_{j=1}^{\#C} \chi^2(v_i, c_i) \tag{2.8}$$

A medida χ^2 identifica o quanto os valores observados nos dados se distanciam do valor esperado. Números altos indicam que a hipótese de independência, também chamada hipótese nula, está incorreta em relação aos dados avaliados (46). Assim, os valores do atributo analisado apresentam impacto na determinação da classe alvo.

A métrica χ^2 oferece vantagens em relação a outra métrica comum, o ganho de informação, pois apresenta valores já normalizados. Além disso, proporciona resultados mais efetivos na redução agressiva de termos, com remoção de até 98% do vocabulário sem perda de acurácia na classificação (45).

2.3.3 F-Medida

F-Medida, também chamada F1, é dada pela média harmônica entre a precisão e a revocação avaliadas nos dados de treino. Sua fórmula é apresentada na equação 2.9 (38).

$$F - Medida = \frac{2 \times Revocação \times Precisão}{Revocação + Precisão} \tag{2.9}$$

As métricas Precisão e Revocação, por sua vez, são apresentadas nas fórmulas 2.10 e 2.11, onde PV define positivos verdadeiros, PF define positivos falsos e NF negativos falsos.

$$Precisão = \frac{PV}{PV + PF} \tag{2.10}$$

$$Revocação = \frac{PV}{PV + NF} \tag{2.11}$$

Segundo (38), uma das motivações para o uso de F1 na seleção de atributos é que esta costuma ser, na maioria dos estudos, a principal métrica aplicada na avaliação dos classificadores, resultando em uma pré otimização dos resultados. O trabalho (38) ressalta também que esta medida favorece a dependência dos atributos à classe positiva em relação à negativa.

2.3.4 Correlação

Dado uma característica x_i do espaço de atributos X ($x_i \in X$) e Y o atributo alvo, a fórmula do coeficiente de correlação de Pearson é mostrada em 2.12 (43).

$$Correlação(x_i, Y) = \frac{covariância(x_i, Y)}{\sqrt{variância(x_i) \times variância(Y)}} \tag{2.12}$$

O coeficiente de correlação apresenta valores entre -1 e 1, sendo que no primeiro caso indica que ambas as variáveis variam juntas mas em direções opostas, e no segundo na mesma direção. Quando o coeficiente se aproxima de 0 aponta que as variáveis não demonstram relação entre si.

Apesar de calcular a co-variação de valores em duas variáveis, o coeficiente não implica em relação causal na forma $x_i \Rightarrow Y$, podendo as oscilações serem motivadas por ainda outras variáveis não abordadas. Outra limitação é que através desta métrica somente relações lineares podem ser avaliadas (47).

O fator de correlação é versátil tanto para a aplicação em dados discretos quanto contínuos, logo pode ser aplicado tanto em problemas de classificação quanto regressão. Esta característica é explorada nos experimentos das seções 5.1.2 e 5.1.3, tirando proveito da classe alvo ser originária de um dado contínuo: o preço.

2.4 Mineração de padrões

A mineração de padrões consiste em analisar um sistema S baseado em suas variáveis, tanto de entrada quanto saída, com o intuito de aproximar seu comportamento através de um modelo computacional. Assumindo que os dados de entrada são representados pelo vetor $\vec{X}$ e o resultado ideal por y, existe uma função objetivo desconhecida $f : \vec{X} \Rightarrow y$ que aproxima fielmente o comportamento do sistema S (34).

No caso dos algoritmos de aprendizado de máquina, utiliza-se um conjunto finito de exemplos de entrada e saída para encontrar, por meio de indução, uma função $g : \vec{X} \Rightarrow y$ entre o conjunto de hipóteses H que aproxime o comportamento da função ideal f (34). A função g é então usada para analisar novos dados e estimar o comportamento do sistema S dadas as novas entradas (48).

Além da dificuldade em explorar o espaço de hipóteses H de forma eficiente, problemas no mundo real costumam envolver muito mais parâmetros do que os observados, como na Figura 5, elevando ainda mais a complexidade em encontrar o modelo ideal (48).

Existem diversas abordagens usadas na obtenção de classificadores, variando entre modelos baseados em distância, probabilísticos, baseados em procura, baseados em otimi-

Figura 5 – Entradas não observadas em um sistema.

Fonte: Traduzido de (48).

zação e ainda baseados em comitê de diferentes técnicas (49). A seguir são apresentadas algumas das técnicas mais difundidas para a busca de modelos de classificação.

2.4.1 K Vizinhos Mais Próximos (*KNN*)

O *KNN* é um algoritmo cujo funcionamento se baseia na projeção dos registros em um espaço vetorial, onde cada atributo é representado por um vértice de um plano N-Dimensional. Após projetar os dados de exemplo (treino), ao tentar prever novos registros estes são mapeados ao mesmo espaço, recuperando os K vizinhos mais próximos, com auxílio de uma metodologia de cálculo distância pré definida. Com base na classe atribuida a maioria dos vizinhos durante o treino, o novo registro tem sua própria classe definida nos testes (48).

Por ser um classificador baseado em exemplos, é capaz de tratar classes não linearmente separáveis. Por outro lado, por ser um modelo não paramétrico, apresenta custo computacional elevado, tanto de processamento quanto de memória, pois quanto maior o dataset de treino, maior a complexidade do classificador (41).

Apesar de ser uma técnica simples, o *KNN* apresenta desempenho notável em boa parte das aplicações, por vezes com desempenho equiparável a técnicas mais robustas como *SVM* e RNA (41, 48).

2.4.2 Regressão Logística

O uso do modelo logístico como classificador toma proveito de uma função sigmoid para, baseado nos atributos fornecidos na entrada, produzir um valor de 0 a 1 que, quando discretizado por um limiar de divisão, proporciona a predição de uma classe binária (34).

O atributo dependente, ou atributo alvo a ser previsto, é modelado como uma função de uma ou mais variáveis independentes. A cada atributo de entrada, que deve ser numérico, é atribuído um coeficiênte, tal que o resultado é dado pela soma ponderada destes aplicada à função sigmoid (50).

2.4.3 Máquina de Vetores de Suporte (*SVM*)

SVM é uma técnica cuja abordagem consiste em maximizar a margem de separação das classes através de vetores auxiliares (ou de suporte) (48).

Considerando como exemplo um problema de classificação binária baseada em dois atributos, após projetar os registros em um espaço bidimensional, é traçada uma reta que não só separe as classes, mas que também produza a maior margem de separação. De maneira genérica, para problemas cujo espaço de atributos é N-Dimensional, tanto a linha de separação quanto os vetores de suporte são descritos por hiperplanos.

Uma das principais vantagens do *SVM*, no entanto, é a possibilidade do uso de funções de kernel para o tratamento de dados não linearmente separáveis. As funções de kernel projetam os dados originais a um espaço de atributos de maior dimensão e, quando bem escolhido o tipo de kernel, tornam mais evidente a fronteira de separação das classes. Esse processo é chamado transformação não-linear (*Nonlinear Transformation*) (34, 48).

Os tipos mais comuns de funcões de kernel são polinomial, hiperbólico e funções de base radial (do inglês *radial basis function*, ou *RBF*, ou simplesmente kernel Gaussiano). A escolha do kernel correto e seus parâmetros ótimos depende muito da característica dos dados, sendo recomendado o uso de busca em *grid* para testar as várias possibilidades.

2.4.4 Árvore de Decisão

Árvores de decisão funcionam baseadas no mapeamento dos valores dos atributos de entrada à regras que, quando encadeadas na forma de caminhos (ou ramos), levam a decisão entre as classes alvo de um problema de classificação (35).

Os caminhos que levam a classificação podem ser vistos como expressões booleanas, tal que, dado o conjunto disjuntos de regras $R = \{r_1, r_2, ..., r_n\}$ de um caminho, a atribuição de uma classe C_k é dada pela expressão $r_1 \wedge r_2 \wedge ... \wedge r_n \Rightarrow C_k$. Uma mesma classe pode ser atribuida por diversos caminhos diferentes, gerando expressões boleanas mais complexas (35).

O encadeamento de regras é otimizado pela escolha de atributos que fornecem o maior ganho de informação, ou entropia, tal que uma maior representatividade possa ser atingida com o menor número de regras possível.

Para ser viável computacionalmente, no entanto, a escolha dos atributos e formação dos nós (regras) é feita de maneira heurística, tal que execuções diferentes podem gerar modelos diferentes para o mesmo dataset, sob as mesmas condições.

O algoritmo *CART*, do inglês *Classification And Regression Trees*, é hoje um dos mais usados para obtenção de modelos baseados em árvore, sendo a implementação adotada pela biblioteca python Scikit Learn. Entre as vantagens desta implementação encontra-se

versatilidade em tratar tanto problemas de regressão quanto classificação, além de ser tolerante a diferentes tipos dados de entrada e valores nulos (51, 35).

2.4.5 Florestas Aleatórias

Ávores de decisão (5.2.2.1) são modelos suscetíveis a sub-ajustamento quando demasiadamente simples (poucas regras, baixa profundidade), e super-ajustamento quando muito complexas, tornando-se um algoritmo de difícil aplicação para problemas mais avançados. Florestas Aleatórias buscam reduzir a variância dos classificadores pela composição de um comitê de diferentes árvores simples, determinando o resultado final por votação (50).

O treinamento individual das árvores é realizado através de técnicas de amostragem, seja no espaço de atributos ou no conjunto de registros de treino (neste caso, também chamados de *Extremely Randomized Trees*) (50). Assim, cada árvore contribui pela captura de características específicas nos dados, compondo juntas a capacidade de predição da classe alvo com separações complexas.

2.4.6 Naïve Bayes

Naïve Bayes possui seu funcionamento semelhante à regressão logística (5.2.1.1), no sentido que também busca efetuar a classificação baseado em uma função de verossimilhança. Contudo, a determinação dos pesos dados a cada atributo é calculada com base em probabilidades condicionais segundo o teorema de Bayes (52, 36).

As probabilidades a *priori* individuais de cada atributo são estimadas da observação das frequências nos registros de treino (52). Como exemplo para um sistema de diagnóstico (classificação binária), a equação 2.13 mostra a probabilidade do registro pertencer à classe positiva C, dados os valores de seus atributos representados pelo vetor $\vec{x} = \langle x_1, ..., x_n \rangle$.

$$p(C|x_1, ..., x_n) = p(C) \prod_{i=1}^{n} p(x_i|C) \tag{2.13}$$

O cáculo das probabilidades a *priori*, no entanto, é condicional ao tipo de domínio dos atributos. Para valores binários (codificados em *one hot encoding*) pode ser usado o modelo de Bernouli, para valores numéricos pertencentes ao conjunto de números naturais o modelo Multinomial é adequado, já para valores numéricos contínuos o modelo Gaussiano é necessário.

2.4.7 Redes Neurais Artificiais: *Multilayer Perceptron*

Redes Neurais Artificiais (RNA) são uma classe de algoritmos inspirados no funcionamento do sistema nervoso, compostos por uma malha de neurônios que se comunicam

através de sinápses (35).

No caso do *Multilayer Perceptron* (*MLP*), a rede é composta por uma camada de entrada, com um neurônio por atributo, uma camada de saída, com um neurônio para problemas de classificação binária ou N neurônios para casos *multilabel* (onde N é a cardinalidade do conjunto de classes), e pelo menos uma camada intermediária.

Cada neurônio recebe impulsos de toda a camada anterior e propaga um novo impulso baseado em sua função de ativação. Os neurônios são calibrados estabelecendo pesos aos estímulos das camadas anteriores, sendo que estes são ajustados por métodos iterativos como *backpropagation* ou gradiente descendente (35).

Quanto maior o número de camadas intermediárias em uma rede, maior a complexidade do modelo e maiores as chances de super ajustamento aos dados de treino.

2.5 Interpretação de modelos

A aplicação de métodos de aprendizagem de máquina no reconhecimento de padrões em sistemas complexos é crescente, contudo ainda há uma lacuna entre a identificação de padrões e extração de conhecimento (53). Apesar da obtenção de modelos preditivos, ou descritivos, ajudar na otimização e automação de diversas tarefas, por vezes a compreenção dos eventos por trás dos padrões são úteis, quando não fundamentais.

A primeira dificuldade na extração de conhecimento de modelos é que o próprio conceito de ”interpretação” pode ser subjetivo (54). Molnar define em (55) a Aprendizagem de Máquina Interpretável como métodos e modelos que fazem as predições e comportamentos de um sistema de aprendizagem de máquina compreensíveis a humanos.

Trabalhos como (54) e (53) tentam organizar o escopo de interpretação de acordo com propósitos diferentes. Em (55) é apresentada uma organização de métodos de análises de maneira mais abrangente e objetiva.

Os tipos de análises podem ser divididos inicialmente em Modelos Intrinsicamente Interpretáveis e Métodos de Análise Post Hoc (55), sendo que no primeiro caso os modelos são avaliados como caixas brancas, internamente, e no segundo como caixas pretas, avaliando seu comportamente através da variação das saídas dadas as entradas.

Quanto à metodologia de intepretação, este trabalho usa o termo ”Interpretação Direta” para análise de Modelos Intrinsicamente Interpretáveis, onde as informações são auferidas diretamente nos classificadores, e ”Interpretação Indireta” para Métodos de Análise Post Hoc, quando é necessário o emprego de *wrappers* na análise.

Outra classificação útil é em escopo de interpretação, local ou global (55). Em (56) esta classificação recebe os nomes de Explanação de Instâncias e Explanação de Modelos, nomenclatura adotada nos capítulos seguintes desta pesquisa.

- **Explanação de modelo**: A lógica do modelo como um todo pode ser analisada a fim de verificar como as decisões são construídas (55). Para um modelo de diagnóstico de doenças baseado em regressão, por exemplo, pode-se verificar se algum dos atributos de entrada é mais decisivo na classificação do que os demais, como a idade ou hábitos alimentares.

- **Explanação de instância**: São analisadas predições individuais, para cada amostra de dado, buscando fundamentar a predição realizada (55). Um exemplo desta abordagem é analisar o caminho de uma árvore de decisão em busca das regras que conduziram a uma predição específica.

Em (55) e (56) o termo "explanação" é adotado para representar elementos que ajudem na interpretação ou compreensão das saídas dos métodos de aprendizagem de máquina. Para facilitar o julgamento das explanações, os autores propõem métricas variadas de avaliação. Para metodologias, (55) aponta as seguintes qualidades:

- Expressividade: Define a qualidade da linguagem ou estrutura das explanações. Regras SE-ENTÃO, por exemplo, são muito mais expressivas do que a malha de neurônios de uma rede neural artificial.

- Transparência: Quão direta é a relação da explanação com o modelo? Modelos intrínsicamente interpretáveis são altamente transparentes, por exemplo.

- Portabilidade: Descreve a especificidade da explanação, se é possível aplicá-la a outros modelos.

- Complexidade Algorítmica: Custo computacional e viabilidade de execução.

Para explanações individuais, avalia-se:

- Confiabilidade: Em (55) e (56) são discutidas a Fidelidade, Certeza, *Novelty* e Representatividade como diferentes abordagens de suporte das explanações, e a Acurácia como grau de generalização. De maneira genérica, estas métricas juntas apontam a confiabilidade da explanação, seu embasamento.

- Consistência: O uso de diferentes técnicas levam à mesma explanação?

- Estabilidade: A variação do mesmo modelo leva a explanações diferentes?

- Grau de importância: Possibilidade de decompor a explanação em elementos de acordo com sua importância.

- Compreensibilidade: Quão legíveis e diretas são as explanação.

2.6 Considerações finais

Os experimentos neste trabalho usam a representação vetorial de texto, com análise multidocumento em nível diário e frequência *TF-IDF*. Para normalização serão adotadas a padronização em caixa baixa, remoção de caracteres especiais e remoção de *stopwords*.

O único tratamento sensível ao idioma, feito especialmente para textos em português, foi a remoção de *stopwords*, com a lista de palavras obtida da biblioteca python *Natural Language Toolkit* (*NLTK*).

Todos os algoritmos na Seção 2.4 serão utilizados em conjunto com a seleção de atributos, como meio de identificar quais as palavras mais relevantes para cada ativo.

3 Trabalhos Relacionados

Estudos indicam resultados promissores na previsão de tendências baseadas na mineração de notícias (8, 9, 57, 10, 7, 58, 59). O tratamento do problema como classificação das variações de preço nas categorias Positiva, Neutra e Negativa é adotado em (8) e (9), sendo que em (57) é considerada apenas a classificação binária nas classes Positivo e Negativo. Trabalhos como (60) e (61) abordam o problema sob a análise de regressão, prevendo não só a direção como a magnitude da variação.

Para classificação, o método para rotular o conjunto de dados de treino em alguns casos é feito manualmente, contudo métodos automatizados baseados nos preços históricos viabilizam amostras de treino maiores, com melhores resultados (8). Por vezes, a rotulação do sentimento nos dados por si só consiste em um problema de pesquisa custoso, como em (62) e (63).

A anotação de notícias segundo a série histórica de preços é apresentada em estudos mais recentes, proposta como alternativa no domínio específico de notícias financeiras, por trabalhos como (64) e (65). A abordagem consiste simplesmente em analisar o comportamento da série histórica e calcular a variação do preço (retorno), tomando como rótulo a direção do preço: queda ou alta (64).

Em (8) e (57) poucos ativos são avaliados, mas ambos aprofundam-se sobre as relações temporais entre o momento da publicação e a reflexão nos preços, relatando resultados melhores quanto menor o atraso. O trabalho descrito em (9) examina uma cesta com 15 ações das quais é possível notar resultados mais significativos para algumas delas em comparação às demais. Além da avaliação individual de ações, pesquisas como (66), (67), (65) e (68) avaliam o humor do mercado acionário como um todo, avaliando índices consolidados como *S&P 500* e *DJIA*, dentre outros.

Apesar de (6) argumentar que ativos financeiros podem permanecer sub ou sobre avaliados por longos períodos, como ocorre em caso de bolhas de mercado, o trabalho (68) argumenta que, a longo prazo, o acumulo de eventos tende a incorporar maior aleatoriedade nos preços, dificultando a identificação de padrões. Em (69) argumenta-se que vinte minutos seria o tempo ótimo de verificação do impacto das publicações, contudo a pesquisa (68) ressalta que mercados emergentes costumam ter menor grau de eficiência.

Em (10) é apresentada uma revisão extensiva de estudos recentes, apontando que a predição do mercado, em especial do mercado de ações e cotação de moedas, é um dos temas principais nos trabalhos de mineração de textos aplicada a finanças. Os autores levantam pontos importantes sobre as pesquisas:

1. Os trabalhos se limitam ao estudo de mercados específicos, com ativos financeiros limitados.

2. O corpo do texto das notícias deve ser preferido ao invés do título, de maneira a reduzir ambiguidades.

O primeiro item deve ser considerado com grande relevância, pois descreve o contexto onde os padrões são observados. O mercado financeiro é adaptativo e padrões dificilmente são permanentes, logo entender os eventos por trás das oscilações fornece indicações se os padrões são pontuais, recorrentes e até generalizáveis para outros contextos.

Uma vez considerados os mercados de acordo com sua eficiência (5), características nos preços negociados na bolsa americana, por exemplo, não necessariamente se aplicam ao comportamento dos preços na bolsa brasileira, em função de suas particularidades. A generalização de padrões obtidos deve ser cautelosa, respeitando variáveis importantes como o mercado observado, o ramo de negócios, o ativo explorado, janelas de tempo e conjuntura política e econômica.

Em (7) foi conduzida uma pesquisa semelhante, ressaltando a complexidade do tópico devido ao seu caráter interdisciplinar envolvendo linguística, finanças comportamentais e apendizado de máquina, como apresentado na Figura 6. Dentre os pontos levantados, muitos corroboram com os apresentados em 10.

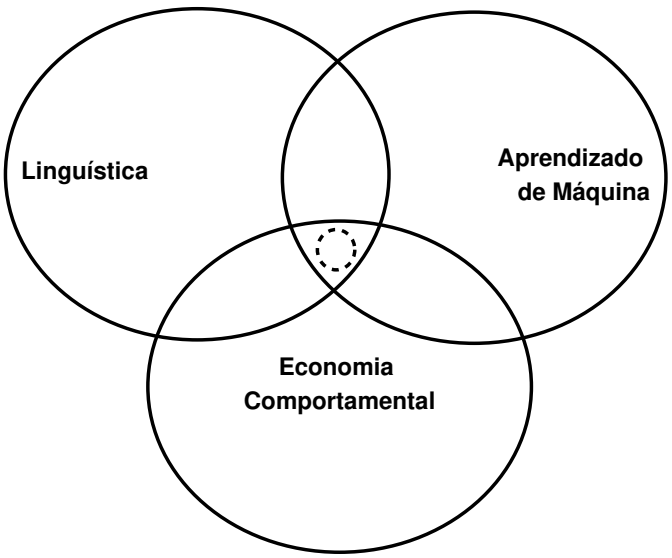


Figura 6 – Dimensões de conhecimentos envolvidos na mineração de texto aplicada ao mercado financeiro.

Fonte: Traduzido de (7).

- A maioria dos estudos trata de problemas de classificação, sendo que as técnicas de Máquina de Vetores de Suporte[1], Naïve Bayes e Redes Neurais Artificiais (RNA)

[1] Mais conhecido pela abreviação em inglês *SVM*, de *Support Vector Machines*

são as mais comuns.

- Medidas de desempenho são relativas e variadas entre os estudos, assim como as bases de dados utilizadas, dificultando comparações diretas dos resultados.

Uma revisão mais recente de artigos é apresentada em (69), reiterando evidências de correlação entre o humor público e as variações de preço. O trabalho apresenta também a discussão sobre anomalias de mercado conhecidas: 1. Efeito Segunda Feira; 2. Efeito Reversa. Estes padrões não só podem ser previstos, como também desenvolvidos com análises computacionais entre dados textuais e movimentações dos mercados financeiros.

Apesar da quantidade de estudos com objetivos semelhantes estar em ascensão, observa-se que pequenas variações na metodologia da pesquisa, tais como a escolha dos ativos, a quantidade de dados, o filtro e tratamento aplicados, métodos e parâmetros de mineração empregados podem impactar drasticamente nos resultados obtidos. Desta forma, novas abordagens devem conduzir a um melhor entendimento do problema bem como indicar os melhores caminhos a se tomar.

Além dos pontos levantados pelos autores, notam-se ainda outras particularidades importantes sobre os artigos que estes abordaram:

1. Como problema de classificação, a medida de desempenho normalmente é elaborada sobre a quantidade de acertos do modelo quanto a direção dos preços.
2. Custos transacionais, taxas e emolumentos são muitas vezes negligenciados na avaliação de viabilidade dos modelos.
3. A maioria das pesquisas usam artigos de notícias ou *twitters*, eventualmente juntos, combinados com dados quantitativos históricos do mercado.

A metodologia de avaliação de desempenho é fundamental no treinamento e seleção de modelos de mineração de dados, contudo em termos de verificação de estratégias de investimento, acertar a direção tomada pelos preços não necessariamente define um bom modelo. Isto pode ser observado, por exemplo, no caso de um modelo que consiga prever consistentemente a direção dos preços por 20 dias, contudo no vigésimo primeiro dia falha ao não antecipar uma queda acentuada, resultando em resultados negativos para todo o período observado. Visto este caso, considerar a taxa de retorno do investimento em um dado período torna-se um objetivo mais interessante do que acertar a direção dos preços.

No contexto brasileiro, no trabalho (4) são debatidos artigos sobre a eficiência na negociacão de ações no Brasil, ressaltando que a maioria deles suporta a eficiência em sua forma fraca. Quando se fala em negociações, esse termo se refere principalmente à capacidade dos agentes envolvidos agirem de forma racional e negociarem ativos a

preços ótimos, sem possibilidade de arbitragem. A eficiência é abordada na literatura econômica como um reflexo do acesso à informação, tal que, quanto mais informados forem os investidores, menores são as chances de se antecipar oportunidades evidentes de lucro (19).

Em (70) é feita uma revisão de dados estatísticos da Bovespa com o intuito de validar a eficiência do mercado acionário no Brasil. Fundamentado sobretudo na análise de multifatores (5), apresentam indícios de que as dinâmicas de negociações se comportam de maneira fracamente eficiente, como já cogitado na pesquisa (4).

O trabalho (19) desenvolve um estudo extenso sobre perfis de investidores, com enfoque no mercado brasileiro, descrevendo como elementos de finanças comportamentais afetam suas escolhas. O autor verifica que investidores se baseiam tanto em acontecimentos passados quanto em expectativas futuras, ressaltando estratégias de investimento preferidas por cada classe de investidor.

Em se tratando de estudos recentes envolvendo métodos automáticos na análise do mercado de ações no Brasil, dentre os trabalhos levantados, o estudo (71) foi o que apresentou uma pesquisa mais abrangente e sistemática sobre o assunto. O objeto de seu estudo foi a ação da petrolífera Petrobrás (PETR4), com emprego de Redes Neurais Artificiais na tentativa de antecipar a direção nas mudanças de preços. Os dados utilizados variam desde indicadores de análise técnica a informações fundamentalistas e macroeconomicas. Os resultados mostram direções previstas com 93,62% de acertos.

Na pesquisa (61) foi avaliada a mesma ação PETR4, contudo sob uma análise de regressão, usando Redes Neurais Artificiais *MLP*. Comparando as predições da rede neural com os valores observados e estimativas obtidas pelo modelo de passeios aleatórios, o autor conclui que o regressor obteve resultados substânciais, antecipando inclusive quedas acentuadas.

Em (72), é proposta a aplicação de dois tipos de Redes Neurais Artificiais, *Multi Layer Perceptron* e Redes Neurais Híbridas, sobre indicadores fundamentalistas na tentativa de antecipar a falência de empresas nacionais. O estudo analisou empresas de diversos setores, argumentando que os modelos são capazes de prever falências com um ano de antecedência.

Em (11), os autores sugerem o uso de conteúdo textual para a identificação de tendências. Em sua metodologia, são coletadas opiniões de diversas fontes, calculando um coeficiente de positividade diário para o conjunto de notícias. Em seguida, o autor avalia a correlação entre os preços negociados para um grupo de ações e o coeficiênte de sentimento expressado nas notícias. Em seus resultados, são apresentados casos relevantes de correlação entre preços praticados e o sentimento nas notícias. O estudo (12) realizou uma pesquisa similar, incorporando elementos semânticos na análise léxica, mas com uma

amostra reduzida de dados.

No trabalho (58) os autores analisam dados textuais de notícias com o objetivo de prever variações, de curto prazo, no desempenho de setores econômicos na bolsa de valores brasileira. São usados como métodos preditivos classificadores baseados em árvore, sendo que os autores optam por traduzir as notícias para o inglês antes fornecê-las aos modelos.

A pesquisa (13) também usa dados textuais, capturados da rede social Twitter, em português, para auxiliar na tomada de decisão para compra e venda de ações na bolsa brasileira. A pesquisa, no entanto, sugere o uso dos dados textuais para verificar previamente o sentimento dos investidores, em seguida fornecendo esta métrica, juntamente com outros indicadores, a estratégias de análise técnica para a tomada de decisão.

3.1 Comparativo de abordagens

Os trabalhos apresentados no capítulo 3 possuem uma característica comum: todos buscam técnicas de análise de métricas financeiras. Por outro lado, o desmembramento dos detalhes na metodologia de pesquisa são bastante variáveis. Para fins comparativos, foram organizadas as pesquisas de acordo com as seguintes perguntas:

Q1. Qual a métrica financeira alvo da análise?

Q2. Qual o mercado financeiro?

Q3. Quantos ativos foram avaliados? Se poucos, quais?

Q4. Quais os dados usados para predição? Se texto, português ou inglês?

Q5. Qual o objetivo?

Q6. Qual a abordagem?

A Tabela 2 apresenta um resumo sobre os trabalhos levantados, organizados por ordem cronológica. Houve a tentativa de organizar as respostas em um domínio comum, contudo o conjunto de respostas não é disjunto, tal que um trabalho pode apresentar mais de uma alternativa para uma mesma questão. Além disso, há casos onde a pesquisa sequer se encaixa no contexto da pergunta. Para os casos onde não foi possível determinar de maneira clara as resposta foi adotado o valor ” – ”.

Quanto à questão Q3, para os casos onde poucos ativos foram avaliados, foi colocado na tabela o nome dos papéis ao invés da quantidade. Em alguns casos não é possível determinar claramente os papéis analisados, pois os trabalhos partem de uma lista inicial mas abordam somente as empresas associadas a eventos identificados. Para estes casos foi usado o valor "Variável".

Tabela 2 – Comparativo de estudos relacionados, com destaque para trabalhos abordando o mercado brasileiro.

Estudo	Ano de Publ.	Q1. Atributo alvo	Q2. Qual mercado	Q3. Quantos ativos	Q4. Dados de treino	Q5. Objetivo	Q6. Abordagem
72	2005	Falência corporativa	Brasil	29	Indicadores Fundamentalistas	Previsão	RNA
11	2008	Cotação ação	Brasil	Variável	Notícias em Português	Avaliação de impacto	Análise de Sentimento
57	2009	Cotação ação	EUA	Variável	Notícias em Inglês	Previsão	Árvore de Decisão, *SVM* e Naïve Bayes
63	2009	Cotação ação	EUA	Variável	Notícias em Inglês	Previsão	Proprietário
8	2011	Cotação ação	Noruega	9	Notícias em Inglês	Previsão	Análise de Sentimento
67	2011	Cotação índice	EUA	DJIA	Twitter em Inglês	Previsão	RNA
66	2012	Cotação índice	EUA	DJIA	Twitter em Inglês	Previsão; Otimização de Portifólio	Regressão, *SVM* & RNA
12	2012	Cotação ação	Brasil	OGX Marfrig	Notícias em Português	Otimização de Portifólio	Análise de Sentimento
65	2013	Cotação índice	EUA	S&P 500	Twitter em Inglês	Previsão	–
68	2013	Cotação índice	Malásia; Tailândia; Indonésia; Filipinas	5	Indicadores Técnicos	Estratégia Automática	Análise Técnica
71	2013	Cotação ação	Brasil	PETR4	Indicadores Técnicos e Fundamentalistas	Previsão	RNA
7	2014	–	–	–	–	*Survey*	–
9	2014	Cotação ação	EUA; Índia	15	Notícias em Inglês	Previsão	Análise de Sentimento
64	2015	Cotação ação	EUA; Alemanha; Inglaterra	Variável	Notícias em Inglês	Previsão	Naïve Bayes
13	2015	Cotação ação	Brasil	9	Twitter em Português e Indicadores Técnicos	Previsão	Análise de Sentimento e Análise Técnica
59	2015	Cotação ação	Hong Kong	22	Notícias em Inglês	Previsão	Sumarização & *SVM*
10	2016	–	–	–	–	*Survey*	–
69	2018	–	–	–	–	*Survey*	–
62	2018	Sentimento no mercado	Arábia Saudita	–	Twitter em Árabe	Rotulação	Naïve Bayes & *SVM*
61	2018	Cotação ação	Brasil	PETR4	Indicadores Técnicos	Regressão	*MLP*
58	2018	Desempenho de setores	Brasil	9	Notícias em Por. traduzidas ao Ing.	Previsão	Proprietário
60	2019	Cotação ação	Índia	6	Twitter em Inglês	Regressão	*KNN*, *SVR* & Alg. Genéticos

Da Tabela 2 é possível perceber que a quantidade de trabalhos aplicando técnicas preditivas ao mercado de ações no Brasil ainda é baixa, sobretudo com o emprego de conteúdo textual em português. A falta de pesquisas associada à baixa quantidade de ativos avaliados ainda leva a dúvidas com relação a viabilidade no tratamento do problema.

Dadas as metodologias adotadas nos estudos avaliados, bem como suas sugestões para trabalhos futuros, os seguintes pontos de melhoria foram adotados para os experimentos nesta dissertação:

- Estender os estudos (11, 58, 61, 13, 71, 12, 72) sobre aplicação de técnicas automáticas na análise do mercado de ações, especificamente para o caso brasileiro e com dados em português.
- Estudo de um conjunto maior de ativos, com as 64 principais ações brasileiras dadas pelo índice Ibovespa. Os resultados serão analisados tanto indivualmente quanto consolidado por setor.
- Emprego de métricas financeiras na avaliação e seleção de classificadores. Estas métricas são aplicadas em conjunto com medidas tradicionais de desempenho.
- Explorar mais técnicas na modelagem dos classificadores e compará-las com estratégias financeiras já difundidas como *baseline*, estabelencedo um comparativo mais abrangente.

Além dos pontos identificados, no alcance dos trabalhos avaliados, os tópicos a seguir são tomados como inexplorados nas pesquisas anteriores:

- Pela teoria econômica de aversão de perdas (25), fundamentada sobre a teoria do prospecto (73), investidores tendem a reagir de maneira mais emocional a riscos do que ganhos. Esta dissertação explora este fator com o tratamento do problema com o foco na atencipação de quedas de preço.
- Utilizar o potencial descritivo dos modelos, buscando padrões interpretáveis.
- Estudo e análise dos meios de comunicação como um todo, com escopo multi-documento, sem seleção prévia de notícias ou extração de eventos para cada papel. Busca-se através da seleção de atributos identificar automaticamente os termos mais pertinentes a cada ação.

A proposta da dissertação é discutida em mais detalhes na seção 3.2.

3.2 Contextualização da proposta

Esta pesquisa trata a previsão de quedas de preço sob a ótica de um problema de classificação, usando tanto o título quanto o corpo de notícias para treino dos modelos, e a direção das variações de preço como atributo alvo a ser previsto.

A hipótese da pesquisa é que, conforme sugerido nos estudos (70) e (4), o mercado de ações brasileiro é fracamente eficiente em termos de informação, logo oportunidades de ganhos podem ser obtidas se as informações públicas forem capturadas e processadas em tempo hábil.

O trabalho supõe também que não só o sucesso de estudos internacionais pode ser alcançado para o mercado nacional, como também que algoritmos mais simples podem atingir resultados razoáveis. Segundo o princípio da Navalha de Occam, soluções mais simples são sempre preferíveis (34) e, neste caso, podem conduzir a classificadores informativos que suportem as predições através de elementos interpretáveis.

Dentre os objetivos principais se encontram:

- Avaliar a viabilidade da predição automática de quedas de preço no mercado de ações brasileiro, com auxílio de técnicas de classificação baseadas em aprendizagem de máquina. Os classificadores são analisados como estratégias de operação e comparados com as técnicas tradicionais *Buy & Hold* e médias móveis exponenciais, a fim de averiguar sua capacidade de reduzir riscos e perdas financeiras.

- Analisar empiricamente indícios do grau de eficiência informacional do mercado de ações brasileiro, levantando o período de notícias mais relevante na predição dos preços futuros, assim como por quanto tempo essas informações permanecem pertinentes na decisão dos investidores.

- Determinar se os modelos preditivos podem fornecer informações racionalizáveis aos investidores, de forma a suportar os prognósticos apresentados. Sobretudo, buscam-se palavras relacionadas a eventos que possam desencadear a desvalorização de papéis, avaliando-se em nível micro por ações individuais, e em nível macro por setor de negócios.

Pontos de melhoria na metodologia de pesquisa foram identificados em relação aos trabalhos relacionados (7, 10), os quais serão abordados durante os experimentos, são eles:

- Emprego das métricas financeiras, como retorno de investimento e análise de perdas, além das medidas tradicionais, como precisão, revocação e F-medida. Estas métricas são usadas tanto como medidas a serem otimizadas como na avaliação dos modelos, buscando-se examinar relações entre as medidas de desempenho e financeiras.

- Avaliação de um conjunto maior de técnicas de aprendizagem de máquina: Naïve Bayes, Máquina de Vetores de Suporte, *Multilayer Perceptron*, *KNN*, Árvores de Decisão, Floresta Aleatória e Regressão Logística.
- Dada a teoria de aversão de perdas e resultados em estudos semelhantes, é enfatizada a predição de perdas e redução de riscos.
- Análise abrangente com 64 papéis diferentes, de diferentes setores de negócios, avaliados sob variadas janelas de tempo, a fim de determinar quais as configurações mais promissoras à predição de preços no mercado nacional.

Os principais desafios são:

- A língua portuguesa apresenta maior complexidade que a inglesa, e bibliotecas de processamento de linguagem natural voltadas a este idioma são mais limitadas.
- Não foi encontrado um corpus público com notícias brasileiras, logo os dados tiveram de ser capturados via *crawler* e passaram por toda etapa de pré processamento para produzir uma base de dados nova.

No capítulo 4 é apresentada a metodologia adotada para a análise experimental, contemplando o escopo, insumos, propostas e limitações.

4 Metodologia e protocolo de experimentos

A metodologia de experimentos inicia em 4.1. A seção trata de como as janelas temporais influenciam no objetivo do classificador, determinando se os padrões podem ser preditivos ou apenas justificativos. Em seguida são abordados as janelas temporais, o momento de tomada de decisão e o prazo de validade desta, definindo o escopo dos experimentos.

As etapas que compõem a pesquisa são resumidas no diagrama da Figura 7. O levantamento e tratamento dos dados usados é apresentado na seção 4.2, discorrendo sobre as particularidades individuais dos dados textuais e de preço, passando pela representação escolhida e discutindo a arquitetura de implementação adotada para consolidá-los.

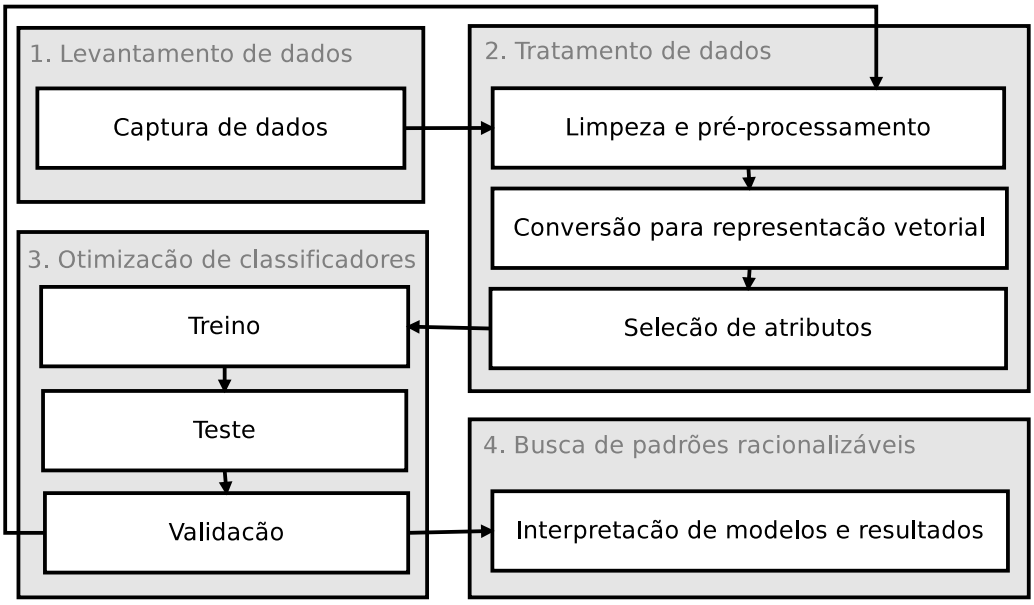


Figura 7 – Fluxograma de etapas para a condução dos experimentos.

Fonte: Produzido pelo autor

Em 4.3 é apresentada a metodologia escolhida para treinamento e teste dos modelos de classificação, discorrendo sobre o número de iterações de validação, treino e teste no processo, os subconjuntos de dados empregados em cada uma destas etapas e a metodologia de amostragem. A seção 4.4 aprofunda-se nas medidas de desempenho aplicadas na avaliação dos classificadores.

Devido à escassez de estudos do mercado brasileiro com notícias em português, o tratamento de dados e ajuste de classificadores foi feito de forma iterativa, buscando as configurações que levam aos melhores modelos de classificação. Os experimentos em si são cobertos no capítulo 5, divididos em duas etapas: Otimização de classificadores (5.1) e

Avaliação de padrões (5.2).

4.1 Definição de intervalos temporais

Esta seção descreve as janelas de tempo analisadas nos experimentos. Em 4.1.1 são discutidas duas propostas de tratamento para o problema, avaliando as vantagens e desvantagens de aplicação de cada uma, bem como sua viabilidade.

Em 4.1.2 são abordados as publicações usados para o treinamento dos classificadores, assim como os prazos de retornos a serem previstos. São descritos também os preços a serem empregados no cálculo de retornos, de abertura e fechamento de mercado, bem como o momento no qual as predições são aplicáveis: intervalo de tomada de decisão.

4.1.1 Abordagem preditiva e justificativa

O tempo das observações, tanto das notícias quanto dos preços, é fundamental na definição do objetivo do classificador. Observa-se duas distinções sutís quanto ao propósito, enunciadas a seguir:

1. Baseado nas notícias de hoje, é possível apontar e justificar quedas nos preços?
2. Baseado nas notícias veiculadas nos dias anteriores, é possível evitar perdas financeiras futuras com modelos que justifiquem suas predições?

Para o primeiro caso explora-se um modelo que possa identificar e racionalizar uma baixa no valor da ação. Dadas as notícias publicadas em D_0, tenta-se determinar se a variação de preço entre a abertura e o fechamento do mercado, no mesmo D_0, foi negativa. Apesar de ajudar a fundamentar a variação, esse modelo não é capaz de predizê-la. Isso se deve ao fato de usar as notícias para justificar variações de preço já ocorridas, de forma que é impossível reverter os resultados.

No segundo caso, almeja-se um modelo tanto justificativo quanto preditivo, que propõe uma decisão de compra na abertura do mercado (em D_0) baseado nas notícias disponíveis até então (até D_{-1}). Os reflexos imediatos no preço proporcionados pela notícia são perdidos, em contrapartida variações nos dias seguintes podem ser antecipadas com maior tempo para a tomada de decisão.

A análise para o primeiro caso é melhor executada quando em tempo real, contudo os dados tanto textuais quanto séries temporais usados na pesquisa não dispõem de granularidade intradiária. Por este motivo, e por proporcionar a análise do retorno financeiro comparativa com demais técnicas tradicionais, a segunda abordagem do problema foi escolhida.

4.1.2 Janelas temporais

Existem dois ajustes com relação ao tempo da observação de padrões: as notícias usadas na predição e os retornos a serem previstos. O primeiro chamamos janela de publicação, o segundo janela de retorno. O objetivo é determinar por quanto tempo uma notícia permanece relevante, escolhendo o intervalo de publicações passadas com maior impacto e os prazos futuros com maior potencial preditivo.

- **Janela de publicação**: É a janela deslizante que define a informação passada disponível para a análise. As predições podem ser feitas baseadas nas notícias do dia anterior, da semana anterior ou do mês anterior. Para o caso de uma semana, por exemplo, para cada dia, as predições são baseadas em todas as notícias publicadas nos últimos 5 dias úteis.
- **Janela de retorno**: Define a informação futura a ser prevista: os retornos financeiros. As janelas de retorno também são deslizantes e iniciadas no dia da análise, com predições de 0 dia (variação entre abertura e fechamento do mercado), 5 (uma semana) e 21 dias úteis (um mês). Para a janela semanal, por exemplo, o modelo prediz se o preço será mais alto ou menor cinco dias úteis à frente, desconsiderando as variações intermediárias.

Nos testes executados, a análise é centrada no dia corrente, chamado D_0, no momento da abertura do mercado. Os retornos são sempre calculados em relação ao futuro, enquanto as informações usadas para treino e predição são sempre passadas, contadas a partir de D_{-1}. A Figura 8 ilustra a relação temporal.

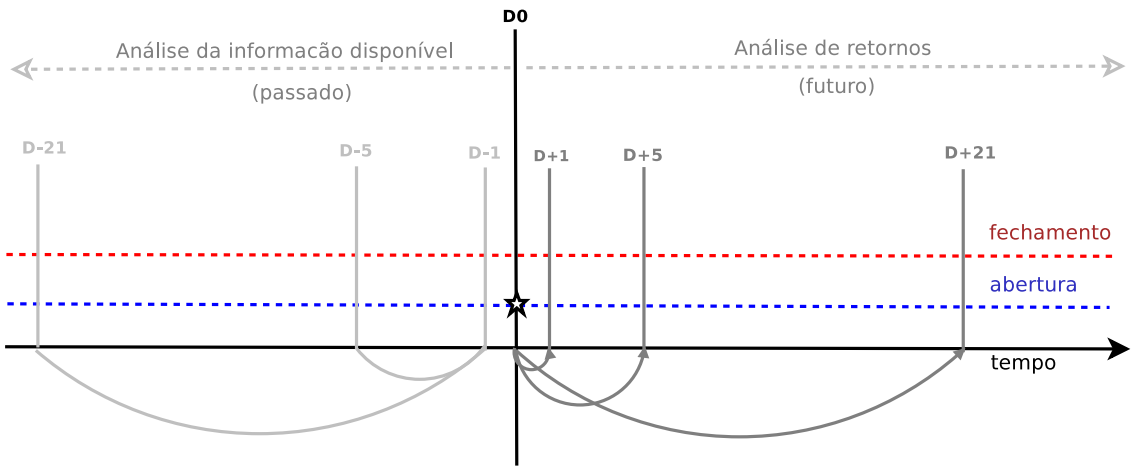


Figura 8 – Janelas de retornos e de publicações.

Os testes foram executados baseados em dias úteis. Notícias publicadas em finais de semana e feriados foram associadas ao dia útil imediatamente anterior. O intervalo de tempo usado no cálculo de retornos também se apresenta sempre em dias úteis, de forma

que um retorno com $\Delta = 1$, quando avaliado em uma sexta-feira, se refere ao preço de segunda em relação ao de sexta, desconsiderando o final de semana.

As janelas de publicação e de retorno são implementadas como janelas deslizantes em relação ao dia analisado (D_0), com tamanho definido por um Δ. Considerando, por exemplo, um artigo publicado em D_{-5} da Figura 8, assumindo-se um intervalo $\Delta = 5$, suas palavras serão consideradas para computar frequências em D_{-5}, D_{-4}, D_{-3}, D_{-2} e D_{-1}.

Para o caso das publicações, o Δ define o conjunto de notícias que serão agrupadas e tratadas como um documento único, seja para o treino ou predição, já para o caso dos retornos, o Δ determina o prazo futuro para o qual direção dos preços (atributo alvo) será prevista.

O agrupamento de notícias foi analisado sob duas abordagens: 1. Pesos uniformes; 2. Ponderação por data de publicação. A ponderação basea-se no cálculo de médias móveis exponênciais com pesos atribuídos segundo a fórmula 4.1, onde N é o Δ total de datas avaliadas e i é quantidade de dias anteriores em relação à data base. O comparativo entre as abordagens é detalhado na seção 5.1.5.

$$w(i) = \begin{cases} i > 0 : w(i) = \dfrac{2}{N+1} * \left[1 - \left(\dfrac{2}{N+1}\right)^{i}\right] \\ i = 0 : w(i) = 1 \end{cases} \tag{4.1}$$

Para o cálculo de retornos busca-se calcular a variação de preço em um dado intervalo de tempo futuro, um delta, tal que se faz necessário determinar qual o período inicial (preço inicial) e o período final (preço final). Como os testes pressupõem a análise em D_0, na abertura do mercado, o preço de abertura em D_0 é assumido como o preço inicial para todos os cálculos de retorno. A motivação baseia-se no fato de este ser o valor da ação imediato no momento da análise.

Tanto para os intervalos de publicação quanto de retorno, foram consideradas inicialmente os períodos de 1, 2, 3, 4, 5, 10 e 21 dias úteis, sendo estes futuros para retornos e passados para publicações. Em análise prévia observou-se que os períodos de 1, 5 e 21 dias eram os mais significativos, pois correspondem respectivamente a um dia, uma semana e um mês em dias úteis. Para retornos, o intervalo de 1 dia foi substituído pelo de 0 dia, representando a diferença entre o preço de abertura e o de fechamento para a mesmo data. Estes períodos foram então adotados como padrões para os testes seguintes.

4.2 Conjunto de dados

Esta seção expõe as características dos dados usados durante os experimentos, desde sua captura, integração e escolha da representação. Em 4.2.1 e 4.2.2 são discutidas as propriedades individuais, respectivamente das notícias e séries de preço.

Em 4.2.3 é apresentada a metodologia pela qual os dados textuais foram modelados para a representação computacional. Discutem-se os desafios no filtro das notícias, o agrupamento dos documentos e escopo da análise, além da métrica escolhida para avaliacão de frequência dos termos.

A seção 4.2.4 aborda, sob o ponto de vista de implementação, como os dados textuais foram associados às séries de preço. A seção discorre ainda sobre a sequência de transformações aplicadas aos dados antes de fornecê-los aos algoritmos de classificação, descrevendo a arquitetura ótima encontrada para facilitar a variação de parâmetros durante os experimentos.

4.2.1 Notícias

Os dados textuais utilizados começaram a ser capturados, automaticamente via *crawler*, a partir de agosto de 2016. Além do texto da notícia em si, foram gravados também meta-dados como a fonte (site de onde o artigo foi obtido), a data da publicação (fornecida pelo site ou a data de captura, quando suprimida), o título e a URL. Os dados foram armazenados em arquivos texto, organizados em subdiretórios por provedor e mês de publicação, para agilizar as buscas.

Os portais avaliados pelo *crawler* são apresentados na Tabela 4. Como, em geral, os provedores publicam notícias com conteúdo de diversos gêneros (esporte, política, entretenimento etc), sempre que possível foram implementados filtros para obter apenas notícias de cunho político ou econômico, principais influenciadores nas decisões dos operadores. A seleção ocorre através da URL do artigo, conforme a segunda coluna da Tabela 4.

Tabela 4 – Portais de notícias avaliados e respectivos filtros.

Provedor	**Filtro de URLs (Regex)**
estadao.com.br	(economia.estadao|politica.estadao)
g1.globo.com.br	(noticia|economia|brasil|politica)
uol.com.br	(economia.uol|noticias.uol)
br.investing.com	(mercado|comunicado|indicadores-econ)
exame.abril.com.br	(exame).abril.com.br
veja.abril.com.br	(veja|vejasp|vejario|exame).abril.com.br
br.adfvn.com	*
infomoney.com.br	*
valor.com.br	*
ultimoinstante.com.br	*

A obtenção dos filtros foi baseada na análise prévia da estrutura de publicação dos provedores, onde foram avaliadas as notícias de maneira geral e então determinados

caminhos de URLs que apresentavam conteúdo mais relevante. Os filtros são compostos de palavras chave aplicadas através de expressões regulares, utilizando o operador "ou" (caractere "|").

O volume mensal de notícias para cada portal é apresentado na Tabela 5. Os dados dos provedores Advfn, Estadão, G1 e UOL foram capturados desde o início do período, enquanto os dados dos portais Investing, Exame, informoney, Último Instante, Valor e Veja foram capturados posteriormente, a partir de julho de 2017, para incrementar o volume diário de artigos.

Tabela 5 – Volume mensal de notícias por portal.

Ano/Mês	Advfn	Investing	Estadão	Exame	G1	Infomoney	Últ. Instante	Uol	Valor	Veja	Total
2016/08	0	0	0	0	59	0	0	8	0	0	67
2016/09	1071	0	980	3	6431	0	0	605	0	4	9094
2016/10	2104	0	1721	1	13420	1	0	1199	0	2	18448
2016/11	2383	0	1655	2	12645	0	0	1252	0	3	17940
2016/12	2331	0	1652	1	10597	0	0	1246	0	7	15834
2017/01	1504	1	1076	7	6950	0	0	895	0	3	10436
2017/02	2354	0	1566	20	9357	1	0	1024	0	15	14337
2017/03	2753	0	1751	21	11004	4	0	1190	1	22	16746
2017/04	2513	0	1669	18	3971	4	0	1209	1	11	9396
2017/05	2822	0	2043	13	2153	6	0	1555	19	12	8623
2017/06	2592	0	1729	32	2039	8	0	1520	37	10	7967
2017/07	2548	7	1525	889	3227	12296	250	1354	2373	2214	26683
2017/08	3163	27	1792	1612	3538	4276	945	1419	4217	3936	24925
2017/09	1921	14	1487	1161	2580	2951	513	1112	3274	2665	17678
2017/10	2462	25	1641	0	3265	3348	788	1300	4092	1937	18858
2017/11	2048	25	1383	0	2733	4535	588	1065	3831	1711	17919
2017/12	1683	19	1152	0	2270	2945	75	944	3121	145	12354
2018/01	437	26	1544	0	2723	3916	679	1208	3551	0	14084
2018/02	0	20	1096	0	1410	2655	513	942	2972	0	9608
2018/03	0	2	123	0	61	735	2	83	470	0	1476
2018/04	0	0	0	0	0	1	0	0	33	0	34
2018/05	0	0	44	0	8	52	0	0	310	0	414
Total	36689	166	27629	3780	100441	37734	4353	21130	28302	12697	272921

É possível notar na Tabela 5 que o volume de notícias é mal distribuído tanto entre os provedores, com G1 representando mais de um terço e Investing menos de 1% dos artigos, e também entre os meses, com volumes muito menores nos meses inciais e finais de captura. A quantidade de notícias por provedor é naturalmente heterogênea, justificada pela diferença de tamanho dos portais, já a diferença entre os meses se devem a problemas no *crawler*, causados por mudanças significativas no código fonte dos provedores.

Apenas notícias cujo acesso é aberto foram apanhadas. Portais como `estadao.com.br` e `g1.com.br` fornecem conteúdo exclusivo para assinantes mediante o pagamento de mensalidades, o que foi desconsiderado durante a pesquisa.

Apesar dos filtros aplicados durante o processo de captura, em análises posteriores

foram encontrados artigos irrelevantes como páginas de erro, *links* em outros idiomas, propaganda e outras páginas que não se referiam à notícia. Em função da quantidade de casos com problemas ser baixa em relação ao corpo completo de notícias, estes foram removidos manualmente.

4.2.2 Séries de preço

Para as séries quantitativas foram avaliados 3 provedores: Exame[1], UOL[2] e Yahoo[3]. Dentre estes, o Yahoo foi escolhido por facilitar a captura automática e por reunir ambos os preços de abertura e fechamento do mercado.

Diferente dos dados textuais, os dados numéricos já são disponibilizados historicamente, tal que a captura foi feita automatizada, mas com execução única. Os atributos extraídos foram os preços de abertura e fechamento, juntamente com o volume de transações, para intervalos diários.

A lista de ações avaliadas compunham a carteira do índice IBovespa no mês de Dezembro de 2018[4]. Os papéis são listados juntamente com a mediana do retornos percentuais para os prazos de 0, 5 e 21 dias na tabela 6. Apesar de haverem ações com claras tendências de alta (MGLU3.SA) e outras de quedas (ELET3.SA), a mediana entre as ações aumenta conforme o prazo.

Consideradas quedas qualquer retorno abaixo de 0, o período apresenta dados razoavelmente balanceados em curto prazo, para a maioria das ações. Para prazos mais distantes, no entanto, a quantidade de quedas ante altas diminui, o que pode ser observado pela mediana na linha total da Tabela 6. Observando a evolução do índice Ibovespa no período na Figura 9, nota-se tendência de elevação na linha de regressão, o que explica a menor quantidade de exemplos de baixa, sobretudo a longo prazo.

As variações na interpretação de quedas são discutida mais a frente nos testes de definição dos rótulos. Em geral, dependendo do *threshold* estabelecido, há variações significativas quanto ao balanceamento entre as classes. O Apêndice A apresenta o balanceamento mais detalhadamente, com os dados usados no experimento da Seção 5.1.6, mostrando a quantidade de amostras de quedas e altas de preço para cada ação, em cada uma das janelas de retorno.

[1] Acessado pelo link: <exame.abril.com.br/mercados/cotacoes>
[2] Acessado pelo link: <cotacoes.economia.uol.com.br>
[3] Acessado pelo link: <finance.yahoo.com>
[4] Acessado pelo link: <http://www.bmfbovespa.com.br/pt_br/produtos/indices/indices-amplos/indice-ibovespa-ibovespa-composicao-da-carteira.htm>

Tabela 6 – Mediana dos retornos percentuais para os prazos de 0d, 5d e 21d, por ação e total.

Ação	**0d**	**5d**	**21d**	**Ação**	**0d**	**5d**	**21d**	**Ação**	**0d**	**5d**	**21d**
ELET3.SA	-0.38	-1.48	-3.35	FLRY3.SA	-0.03	0.44	0.51	GOAU4.SA	0.00	0.83	3.80
CSNA3.SA	-0.35	-0.37	-0.62	WEGE3.SA	-0.02	0.34	2.54	VALE3.SA	0.00	1.01	4.60
ELET6.SA	-0.28	-1.15	-2.16	PCAR4.SA	-0.00	0.82	1.98	FIBR3.SA	0.00	1.37	5.11
GOLL4.SA	-0.27	0.64	8.02	CYRE3.SA	0.00	0.37	0.42	VIVT4.SA	0.00	0.07	-0.61
BRFS3.SA	-0.26	-1.29	-4.52	RAIL3.SA	0.00	1.46	2.35	EQTL3.SA	0.00	0.39	1.28
USIM5.SA	-0.24	0.65	2.27	SUZB3.SA	0.00	2.47	10.44	ENBR3.SA	0.00	-0.10	0.00
CMIG4.SA	-0.23	-0.49	-2.21	QUAL3.SA	0.00	-0.10	0.63	BTOW3.SA	0.00	1.26	4.28
JBSS3.SA	-0.22	0.09	-0.10	PETR4.SA	0.00	1.08	2.47	SBSP3.SA	0.00	0.39	1.08
BRML3.SA	-0.16	-0.18	-1.65	PETR3.SA	0.00	0.89	1.62	CSAN3.SA	0.01	0.40	-0.35
CIEL3.SA	-0.14	-0.95	-2.94	NATU3.SA	0.00	0.29	1.49	BBDC4.SA	0.01	0.75	1.06
GGBR4.SA	-0.12	0.91	2.68	ECOR3.SA	0.00	0.47	1.90	MULT3.SA	0.02	0.00	-0.61
EMBR3.SA	-0.12	0.28	1.94	MRFG3.SA	0.00	-0.14	1.00	ABEV3.SA	0.02	0.19	1.54
LAME4.SA	-0.11	0.12	1.26	TIMP3.SA	0.00	0.82	2.60	SANB11.SA	0.04	0.74	3.15
CPLE6.SA	-0.11	-0.69	-1.04	LREN3.SA	0.00	0.39	1.52	RENT3.SA	0.04	0.98	3.85
EGIE3.SA	-0.11	-0.27	-0.56	SMLS3.SA	0.00	0.80	1.58	B3SA3.SA	0.05	0.66	1.32
BRKM5.SA	-0.09	0.65	2.41	BRAP4.SA	0.00	1.24	4.24	BBAS3.SA	0.05	0.63	2.40
MRVE3.SA	-0.07	-0.16	0.07	UGPA3.SA	0.00	-0.07	-0.47	CVCB3.SA	0.05	1.14	4.40
KROT3.SA	-0.07	-0.63	-1.83	ITSA4.SA	0.00	0.62	1.19	ITUB4.SA	0.08	0.52	1.21
BBSE3.SA	-0.07	-0.31	-0.68	IGTA3.SA	0.00	0.27	0.64	BBDC3.SA	0.08	0.31	0.18
CCRO3.SA	-0.06	-0.70	-2.20	HYPE3.SA	0.00	0.17	1.29	ESTC3.SA	0.18	0.48	2.16
RADL3.SA	-0.04	-0.15	0.58	VVAR3.SA	0.00	0.36	1.98	MGLU3.SA	0.29	3.40	12.63
								Mediana	**0.00**	**0.39**	**1.28**

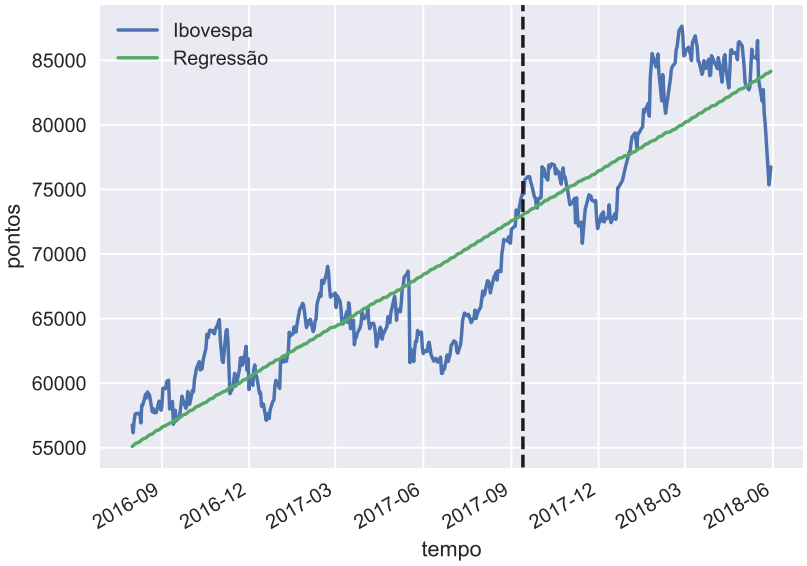


Figura 9 – Índice Ibovespa e regressão linear apresentando tendência. A linha preta divide a série entre os períodos usados no treino e no teste.

4.2.3 Representacão

O trabalho foca-se na análise da frequência de palavras em nível diário. Estabelecido inicialmente o vocabulário, representando o espaço de atributos, é computada a métrica de frequência para cada um dos termos, sumarizadas por intervalos diários. Em outras palavras, todas as notícias de um mesmo dia, ou um intervalo de dias, são tokenizadas e tratadas como um único documento, gerando um registro único para cada data. A Figura 10 exemplifica graficamente a representação adotada.

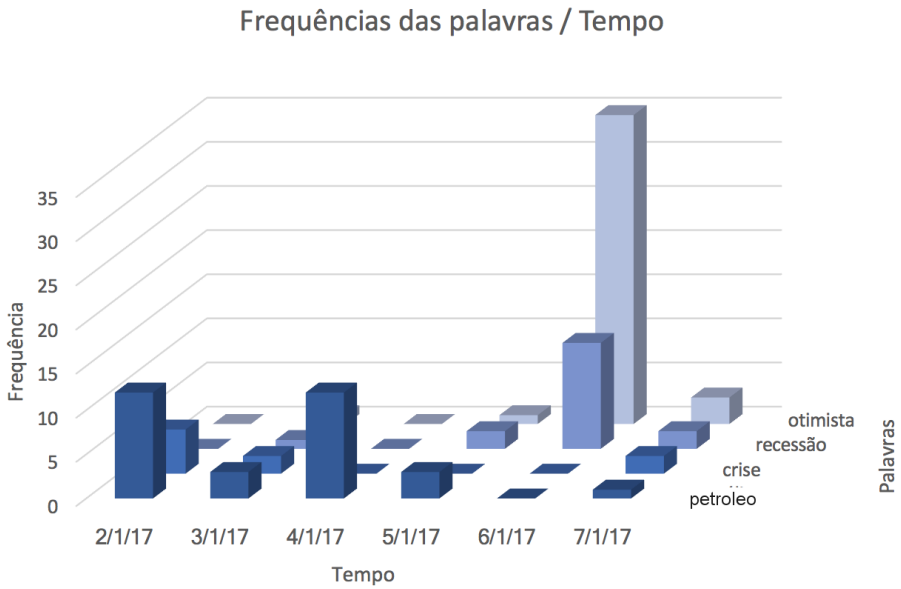


Figura 10 – Exemplo de representação vetorial da frequência de termos em intervalos diários.

Foi cogitada a filtragem de notícias baseada em palavras chaves pré-estabelecidas para cada ação, mas esta metodologia apresentava dois problemas: relatividade na escolha das palavras chaves e redução significativa dos dados. Optou-se então por considerar todas as notícias em uma análise diária dos meios de comunicação como um todo, delegando aos classificadores a tarefa de avaliar a pertinência dos dados. A abrangência das notícias genéricas aumenta a complexidade dos experimentos, entretanto permite a análise de temas ligados à ação tanto direta como indiretamente.

A representação computacional do texto escolhida foi em espaço vetorial, com a frequência das palavras calculada pela metologia *TF-IDF*. Foi considerado o uso de outras métricas de ocorrência de termos como *BoW* e *TF*, mas pelos melhores resultados em testes prévios e em trabalhos relacionados, o *TF-IDF* foi adotado.

Os dados textuais usados foram ambos o título e o corpo da notícia, tratados como texto único. Como pré-processamento do vocabulário foi feita a remoção de *stopwords*, padronização em caixa baixa e remoção de acentuação e caracteres especiais.

A extração de radicais (*stemming*) foi desconsiderada pela possibilidade de prejudi-

car a interpretação dos padrões, ainda assim, foi testada a fim de avaliar se seu impacto poderia ser significativo.

Como seleção de atributos inicial, foi aplicada a redução de dimensionalidade por baixa variância, descartando *tokens* cuja métrica de frequência era constante. Após a redução o espaço de atributos totalizou 327623 termos.

As séries de preço foram usadas sem ajustes, conforme disponibilizadas nos provedores, com os retornos calculados sobre o preço de fechamento em relação ao de abertura. Diferentes metologias foram aplicadas na conversão dos valores numéricos para literais, para serem então empregadas como classes alvo na classificação. Estas metologias são discutidas nas seção 5.1.1.

4.2.4 Implementação do fluxo de dados

Com os dados já capturados, foi desenvolvida a solução para integração das informações e execução dos experimentos em linguagem Python. Foi usada a biblioteca *Natural Language Toolkit* (*NLTK*) para auxiliar no processamento de linguagem natural, Pandas para a manipulação de dados, *Scikit Learn* para treinamento de modelos e Matplotlib e WordCloud para a elaboração de gráficos.

Para facilitar a variação de parâmetros nos experimentos, foi projetada a estrutura de classes conforme a Figura 11. Em resumo os pacotes em cinza (`SerieLoaders`, `Returns` e `Returns2Labels`) foram projetados para fornecerem classes para o tratamento dos dados históricos de preço. Os pacotes em amarelo (`ArticlesFilter`, `ArticlesOrganizer`, `Preprocessing` e `Representation`) tratam do processamento e organização dos dados textuais. O pacote azul (`AttributeSelection`) reúne as classes aplicáveis na seleção de atributos. A classe em vermelho, `Dataset`, é responsável por organizar e garantir consistência das informações para submetê-las aos algoritmos de modelagem de classificadores.

Em termos de dados numéricos, estes são inicialmente parseados por uma das classes do pacote `SerieLoader`, dependendo da fonte de dados. Em seguida, as séries históricas de preço são submetidas a um dos procedimentos de cálculo de retornos disponível no pacote `Returns`. Por último, os retornos são convertidos para uma forma categórica, com uma das classes do pacote `Return2Labels`, para serem usados então como atributo alvo de classificação.

Para os dados textuais, inicialmente é recuperada a lista de arquivos, cada um representando uma notícia, através da classe `ArticlesFilter`. Esta classe é capaz de filtrar as notícias por provedor e palavras chaves, tanto positivas quanto negativas. Para os experimentos desta dissertação foi aplicado apenas o filtro por provedores nacionais.

A lista de arquivos é posteriormente agrupada de acordo com os intervalos de publicação, podendo ser aplicado o agrupamento por dias regulares (`GroupByDays`), dias

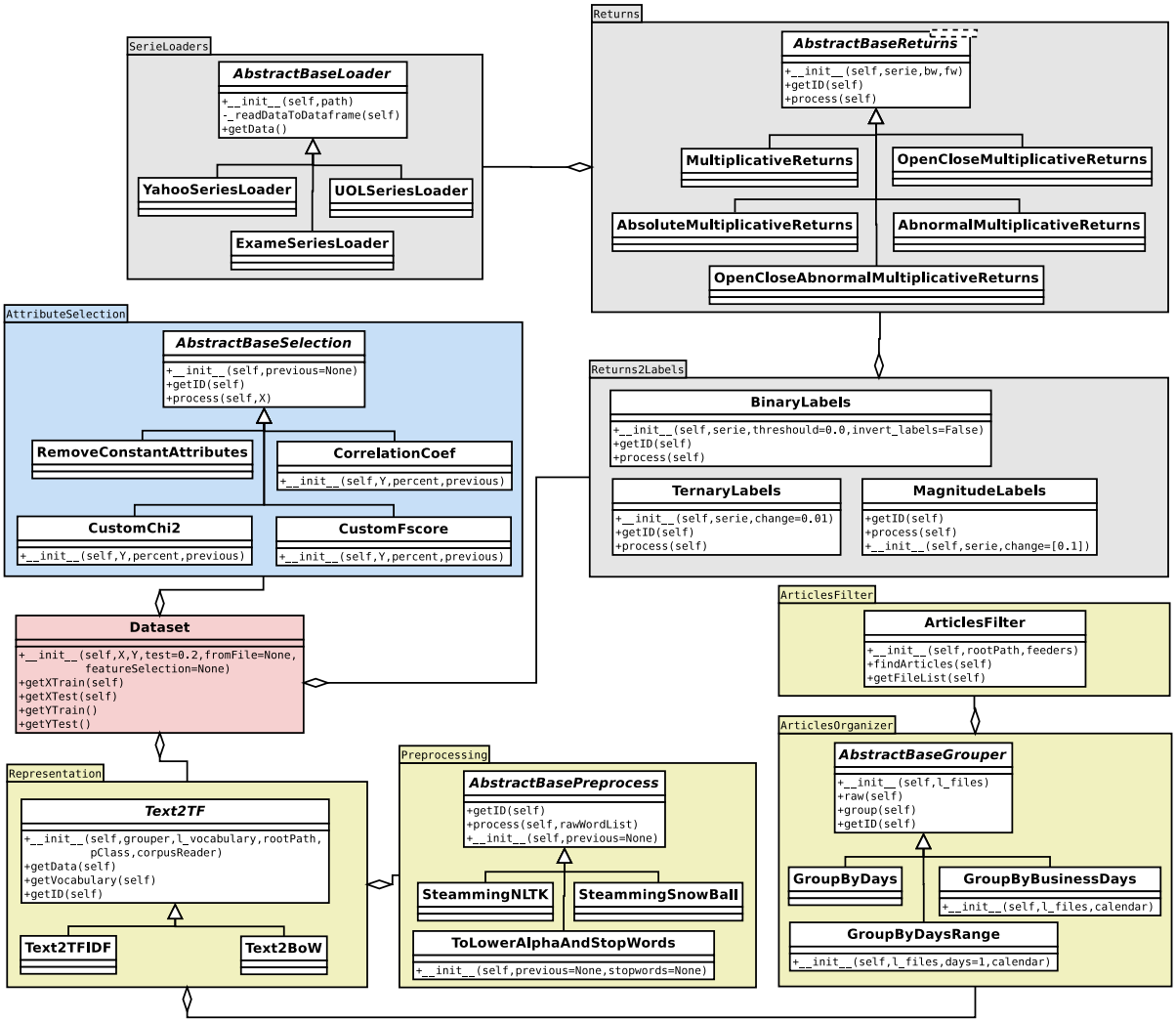


Figura 11 – Diagrama de classes para a modelagem dos dados.

úteis (`GroupByBusinessDays`) ou intervalo de publicação (`GroupByDaysRange`), tanto em dias regulares ou úteis. Estas classes são usadas para ajustar as publicações usadas para o treinamento e predição dos classificadores no capítulo 5.

Por último, a lista de arquivos organizada por data é aplicada a uma das classes de representação do pacote `Representation`, sendo estas responsáveis pela segmentação do texto e avaliação da frequência dos termos. Durante este processo pode ser fornecida uma classe de pré-processamento do vocabulário disponível no pacote `Preprocessing`, com a possibilidade inclusive de aninhar processamentos em sequência.

A classe `Dataset` tem como principal função garantir que as datas tanto nos dados texto quanto numéricos estejam em acordo, descartando pontos onde apenas um dos tipos de dados existe. Esta classe também é responsável por garantir que a divisão entre os dados de treino e teste respeitem a dependência temporal, usando para testes apenas os dados mais recentes. Como a classe `Dataset` já reúne tanto os dados de treino quanto o atributo alvo, pode ser escolhida uma das classes de seleção de atributos no pacote

`AttributeSelection`, também com a possibilidade de aninhar seleções em sequência.

4.2.5 Exemplo do tratamento de dados

As Figuras 12 e 13 mostram o exemplo de duas notícias capturadas do portal Estadão, nos dias 25 e 26 de janeiro de 2018. A primeira linha dos arquivos exibe a URL pela qual o conteúdo foi capturado, a segunda indica o título e as demais o corpo das notícias.

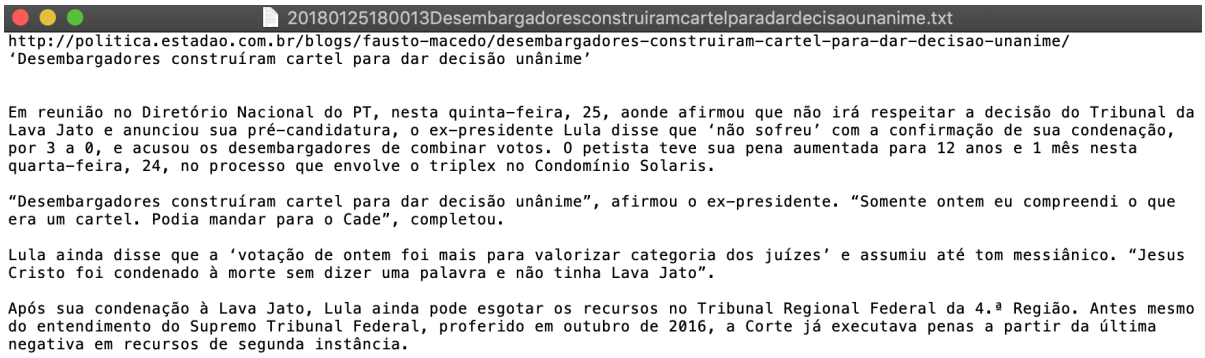


Figura 12 – Exemplo de notícia capturada do portal estadao.com, na data de 2018-01-25.

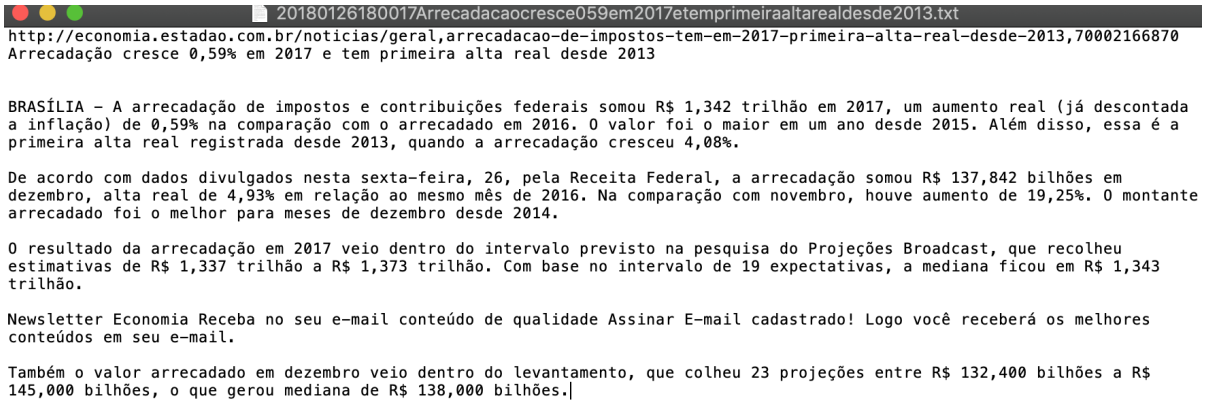


Figura 13 – Exemplo de notícia capturada do portal estadao.com, na data de 2018-01-26.

Após passar pelo agrupamento por data, segmentação de termos e normalizações léxicas (conversão para caixa baixa, remoção de caracteres especiais e remoção de *stopwords*), são calculadas as frequências *TF-IDF* para cada uma das palavras presentes no texto, resultando nos dados apresentados na Figura 14.

A primeira linha na Figura 14 define o espaço de atributos, o vocabulário. A segunda e a terceira mostram as frequências computadas para cada termo. Para este exemplo há apenas uma notícia por data, para os experimentos da Seção 5.1 todas as notícias publicadas em uma mesmo dia são agrupadas, resultando, ainda assim, em apenas um registro por data.

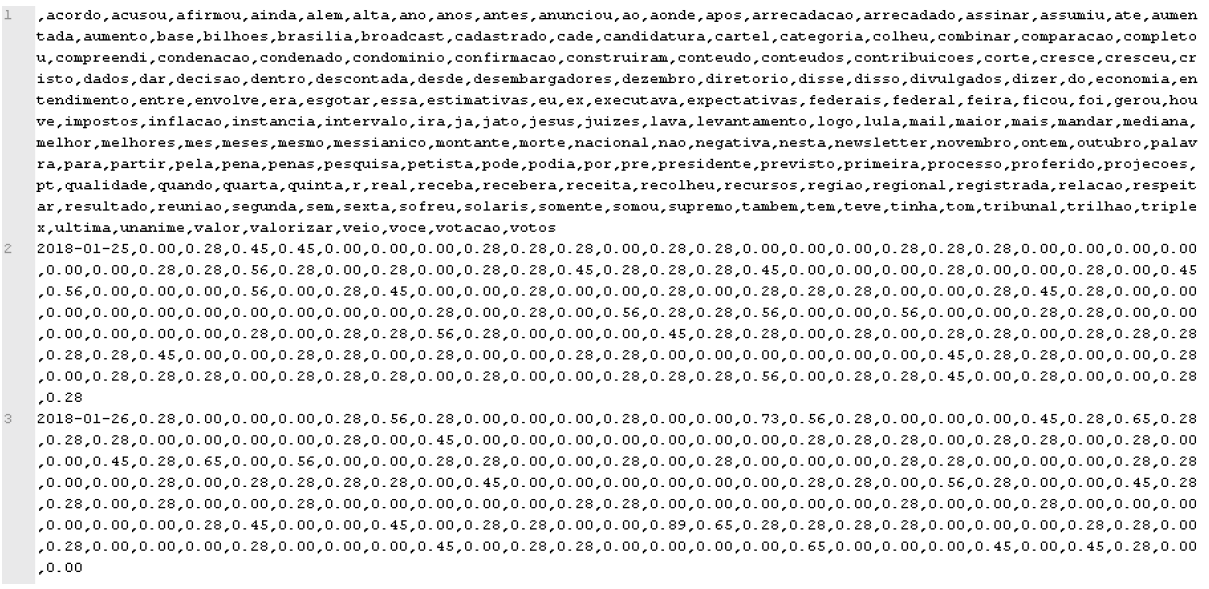


Figura 14 – Exemplo de representação para dados reduzidos, com apenas duas notícias.

Com a frequência dos termos já organizada diariamente, o próximo passo é obter o atributo alvo: a direção dos preços. Para este exemplo foi escolhido o ativo BBAS3.SA, da empresa estatal Banco do Brasil. Os dados referentes aos preços, retornos e rótulos atribuídos a cada dia são mostrados na Tabela 7. O método de rotulação adotado no exemplo é binário, com *threshold* em 0.

Tabela 7 – Dados do ativo BBAS3.SA, para as datas 25 e 26 de janeiro de 2018. A tabela mostra desde o preço de fechamento verificado, passando pelos retornos e pela conversão para as classes de Queda/Alta.

Data	**Preço**	**Fat. Retorno (D+1)**	**Retorno (D+1)**	**Queda=1/Alta=0**
2018-01-25	37.0000	1,0311	0,0311	0
2018-01-26	38.1500	0,9945	-0,0055	1
2018-01-29	37.9400	–	–	–

Enquanto a Figura 14 reúne os dados textuais, convertidos para representação vetorial e organizados diariamente, a coluna “Queda=1/Alta=0” da Tabela 7 define a variação de preço do mercado no período seguinte: Queda ou Alta. A partir destas informações os modelos são treinados com dados históricos, e posteriormente aplicados na predição para datas futuras, dadas as novas frequências *TF-IDF* para os termos.

Este exemplo mostra apenas um caso onde a janela de publicação é de D_{-1} e o de retorno D_{+1}, apenas para a ação BBAS3.SA. Para os experimentos são gerados um conjunto de dados para cada combinação ativo analisado, janela de publicação e janela de retorno avaliado.

4.3 Treino e teste

Tanto para as notícias quanto séries de preço, há dependência temporal entre os dados de treino e teste, pois eventos passados podem influenciar o futuro, mas o contrário não é valido. Foi considerado o uso de técnicas tradicionais como Validação Cruzada e *K-Fold*, contudo as amostragens aleatórias ou estratificadas de registros não são adequadas aos dados temporais por pressuporem independência.

O método *Hold-Out* sugere uma maneira mais simples de separar dados de treino e validação: divide-se os dados em apenas dois grupos, mantendo um percentual de amostras para treino e outro para validação (33). Para os dados temporais, mantém-se sempre a porção inicial dos dados para treino e as datas mais recentes para testes. O treino dos classificadores foi executado em modo *batch*, onde o modelo é treinado apenas uma vez e aplicado na predicão de novas amostras.

Foram reservados 2/3 dos dados para treino e 1/3 para testes, mais precisamente o período compreendido entre 2016/08/01 e 2017/09/12 foi usado para treino, enquanto as datas entre 2017/09/13 e 2018/05/07 foram aplicadas aos testes.

Além da separação entre treino e teste, para os casos de otimização de modelo através de busca em *grid*, foi separada uma parte de validação entre a porção de treino. Os registros de validação são usados para escolher, dentre os parâmetros do classificador, quais produzem os melhores resultados.

As iterações de validação são executadas em subconjuntos do *dataset* conforme na validação cruzada, contudo a divisão entre treino e validação também respeita a dependência temporal. Para tanto, os registros de treino são sucessivamente subdivididos em 1/3 de validação e 2/3 de treino.

Como resultado, subgrupos crescentes de registros são avaliados em cada iteração, sendo que a última iteração roda sobre todos os dados. O funcionamento é ilustrado na Figura 15, onde as duas primeiras barras representam iterações de validação e a última o treino final e teste do classificador.

Como a busca em *grid* de parâmetros demanda longos períodos de execução, para os testes foram usadas duas iterações de validação, a fim de viabilizar a execução de mais experimentos. Mais detalhes sobre o funcionamento das iterações de validação podem ser obtidos no manual da biblioteca Scikit Learn [5].

[5] Acessado em: <scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html>

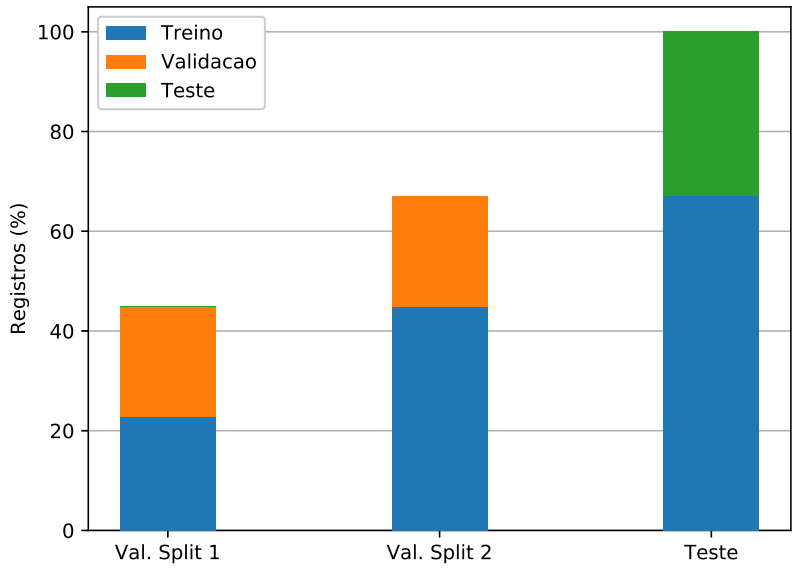


Figura 15 – Validação em dados temporais, exemplo com dois *splits*. Adaptado do manual da biblioteca Scikit Learn.

4.4 Medidas de avaliação

As medidas de avaliação são usadas para mensurar a qualidade do modelo, aplicadas sempre sobre os dados de teste ou validação. A tabela 8 resume as métricas utilizadas e suas fórmulas. O retorno e as perdas financeiras são computadas também para as estratégias Média Móvel Exponencial (MME) e *Buy & Hold*, com o intuito de avaliar a viabilidade do modelo ante as técnicas já difundidas.

Tabela 8 – Medidas de avaliação e suas fórmulas.

Métrica	**Fórmula**	**Descrição**
Revocação	$Revocação = \frac{PV}{PV+NF}$	Onde PV = Positivos Verdadeiros e NF= Negativos Falsos
Precisão	$Precisão = \frac{PV}{PV+PF}$	Onde PV = Positivos Verdadeiros e PF= Positivos Falsos
F-Medida	$F\text{-}Medida = 2 * \frac{Precisão * Revocação}{Precisão + Revocação}$	Médias harmônica entre a precisão e a revocação
Retorno financeiro	$R_F(\Delta_t) = \prod_{i=1}^{n} R_F(\Delta_i)$	Onde $R_F(\Delta_i)$ são os retornos diários da estratégia em cada um dos n dias.
Perda financeira	$P_F(\Delta_t) = \prod_{i=1}^{n} P_F(\Delta_i)$	Onde $P_F(\Delta_i)$ denotam apenas as perdas financeiras sofridas. Em dias onde os resultados foram positivos, $P_F(\Delta_i)$ assume o valor neutro 1.
Perda percentual	$P_P(\Delta_t) = \frac{P_F(\Delta_t)}{P_T(\Delta_t)}$	Onde $P_T(\Delta_t)$ representa todas as perdas ocorridas no período, sejam elas sofridas ou não.

Pela proposta do trabalho, prever corretamente quando ocorrerá uma queda é mais importante do que antecipar cenários de alta. Por outro lado, buscam-se classificadores

capazes de separar entre cenários de queda e de alta de forma eficiente, para que o ganho financeiro das estratégias de operação não seja comprometido.

As medidas *Revocação*, que retrata quantas quedas foram previstas dentre as ocorridas, *Precisão*, que define a quantidade das previsões de queda realmente ocorreram, e F-Medida foram escolhidas como métricas de desempenho. A métrica F-Medida é usada com peso igual tanto para Revocação quanto para Precisão, na forma de média harmônica. Além de promover o equilíbrio entre ganhos e perdas, essa escolha favorece o descarte de classificadores constantes, onde a Revocação é maximizada pela atribuição indiscriminada da classe alvo para todos os exemplos.

Além das medidas de acurácia, métricas financeiras também foram empregadas, avaliadas através de acúmulos diários, com fatores obtidos entre o preço de abertura e o de fechamento, são elas:

1. **Fator de retorno financeiro**: Também chamado de fator de retorno monetário, representa quanto do valor inicial investido se obteve ao final do período de testes. Quando menor que 1, sinaliza perda de patrimônio, quando maior que 1 indica ganhos.

2. **Fator de perda financeira**: Também chamada de fator de perda monetária, esta métrica avalia a perda acumulada ocasionada pelas quedas de preço não previstas, desconsiderando possíveis ganhos no mesmo intervalo. Seu valor também é um percentual do montante inicialmente investido.

3. **Perda percentual**: Semelhante ao fator de perda financeira, contudo essa medida é relativa ao total de perda no período e não ao capital investido inicialmente. Como exemplo, um classificador pode atingir uma perda financeira de 30% no período de testes, porém assumindo que o total no mesmo período foi 80%, o modelo sofreu uma perda percentual de 37.5%.

Ao fazer predições sobre os dados de teste, o classificador sugere os dias do intervalo onde quedas de preço são esperadas. Para estes dias, a posição vendida é considerada no cálculo do retorno financeiro, tal que a variação do dia é descartada no acúmulo dos retornos. De forma oposta, para efeitos de avaliação e comparação, todos os dias não apontados como quedas para o classificador são considerados posição comprada, portanto têm seus retornos relevados durante o acúmulo.

Para testes onde o classificador prevê posições mais longas, como 5 e 21 dias úteis, assume-se que os prognósticos positivos (de quedas) ocorridos anteriormente sobrepõem as predições do intervalo futuro. Este comportamento é exemplificado na Tabela 9.

Tabela 9 – Previsões de longo prazo, um exemplo de 5 dias.

Data	**Fator de Retorno**	**Previsão**	**Posição**
2017-09-19	0.98	0	C
2017-09-20	0.97	1	V
2017-09-21	0.99	0	V
2017-09-22	1.01	1	V
2017-09-25	0.95	1	V
2017-09-26	0.97	0	V
2017-09-27	1.03	0	V
2017-09-28	1.00	0	V
2017-09-29	1.04	0	V
2017-10-02	1.03	0	C
2017-10-02	1.01	0	C

Na Tabela 9, a coluna "Previsão" apresenta o valor 1 para sinalizações positivas de quedas e 0 para negativas. A coluna "Posição" mostra o valor C para posições compradas e V para vendidas, baseadas na sugestão do classificador.

A data 2017/09/19 mantém a posição C em acordo com a previsão de queda do classificador para o próprio dia, pois não houve indicações prévias. Já as datas 2017/09/21, 2017/09/26, 2017/09/27, 2017/09/28 e 2017/09/29 assumem a posição vendida V por haverem previsões anteriores, dentro do intervalo de análise, indicando que a posição deve ser mantida. O fator de retorno financeiro do classificador exposto na Tabela 9 é dado por $(0.98 * 1.03 * 1.01) = 1,019$, enquanto a perda é dada pelo fator $(1 - 0.98) = 0.02$, pois a única perda não antecipada foi em 2017-09-19.

A perda percentual foi usada na comparação de modelos treinados para diferentes ações, reduzindo o viés da análise. Assim, classificadores de papéis com maiores quedas no período não são penalizados ante os demais.

4.5 Descrição do método

A Figura 16 mostra as etapas que compõem a busca pelos modelos preditivos. A primeira etapa consiste em selecionar, entre as 64 ações analisadas, o ativo alvo da predição, representado por s. Em sequência são definidas a janela de publicações, janela de retornos e tipo de ponderação de notícias, representados respectivamente por k, p e w. Estas informações definem o objetivo do modelo, com a ação e prazos a serem previstos, e as informações onde o modelo executará o aprendizado, com as publicações e sua ponderação.

O passo três consiste em agrupar as notícias e indexá-las por data. Dependendo da janela de publicação escolhida, serão agrupadas na mesma data notícias do último dia (D_{-1}), última semana (D_{-5}) ou último mês (D_{-21}), com janelas deslizantes que permitem

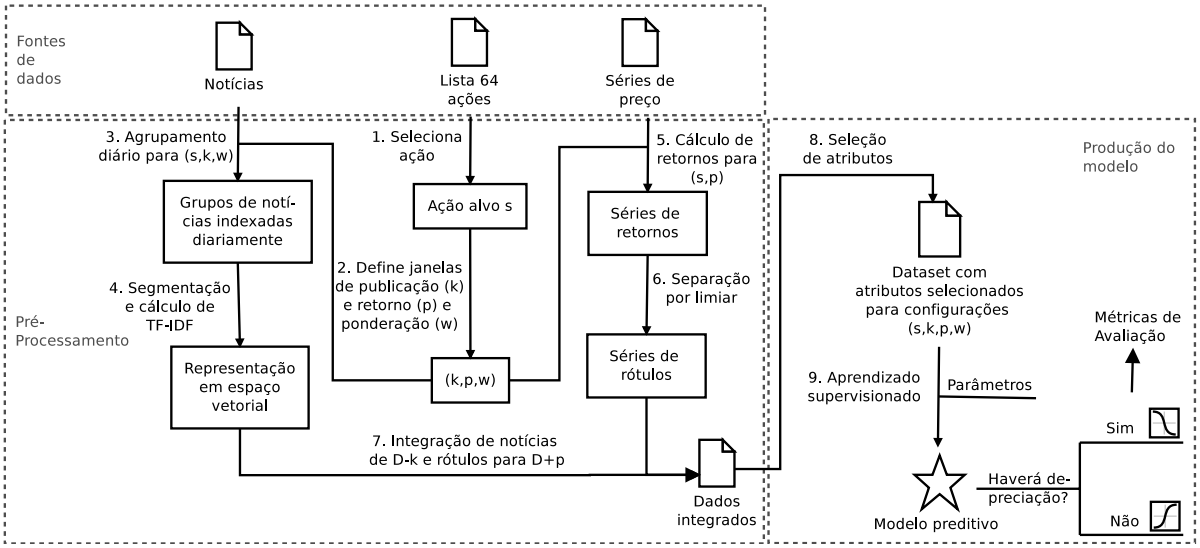


Figura 16 – Transformação das notícias para dados de treinamento de modelos.

que uma notícia afete as frequências do vocabulário em dias subsequentes.

O quarto passo trata todas notícias indexadas em uma mesma data como um documento único, segmentando suas palavras e computando a frequência associadas ao dia. O processo se repete, gerando uma amostra por dia e produzindo ao final a representação em espaço vetorial do *corpus*.

Dada a ação s e a janela de retorno p, as etapas 5 e 6 calculam o retorno diário e os transformam em rótulos. O retorno se dá pelo quociente entre o preço de fechamento em D_{+p} e o preço de abertura em D_0. A transformação dos retornos em rótulos é feita por *threshold*, estabelecendo rótulos diferentes dependendo da faixa de retorno.

No passo 7 as linhas da representação em espaço vetorial são integradas à série de rótulos, pela data, tal que cada linha do *dataset* integrado apresenta as notícias indexadas para D_{-k} e os retornos no prazo futuro, para D_{+p}. O *dataset* produzido possui alta dimensionalidade devido ao grande volume de notícias.

No passo 8 a seleção de atributos é aplicada comparando a frequência das palavras com os rótulos, analisando assim a relação dos termos com a ação. Como resultado, o *dataset* é reduzido a um subconjunto de atributos com as palavras do vocabulário mais relacionadas ao ativo s, para o prazo p, segundo o filtro.

Por fim, no passo 9, as amostras são divididas entre os conjuntos de treino e teste e aplicadas na produção do modelo. O conjunto de treino é aplicado ao aprendizado supervisionado através das técnicas descritas na Seção 2.4, o conjunto de testes é usado para produzir métricas de avaliação da qualidade do modelo descritas na Seção 4.4.

4.6 Considerações finais

Os experimentos serão voltados à análise preditiva, com treinamento de classificadores baseado em dados de publicações de 1, 5 e 21 dias úteis atrás, mapeadas para representação vetorial com frequência avaliada via métrica *TF-IDF*. O atributo alvo a ser previsto é a direção dos preços, obtida através da discretização do retorno financeiro.

Como transformação padrão nos dados textuais são aplicadas a padronização em caixa baixa, remoção de caracteres especiais e remoção de *stopwords*. A seleção de atributos por variância $Var(X) \neq 0$ também foi aplicada como padrão, removendo atributos cujos valores são constantes.

São treinados classificadores para a predição da direção dos retornos para os prazos de 0 (variação entre os preços de abertura e fechamento para o mesmo dia), 5 e 21 dias úteis. A avaliação de desempenho é efetuada através das medidas de precisão, revocação e F-Medida, juntamente com as métricas de retorno e perda financeira e perda percentual.

Serão reservados 2/3 dos dados para treinamento e validação dos modelos e 1/3 para teste, respeitando a dependência temporal. Os algoritmos usados, variações no tratamento das publicações, métodos de conversão das séries de preço para rótulos e detalhes sobre a seleção de atributos são discutidos no capítulo 5.

As limitações dos experimentos são:

- O escopo de investigação é dado em um período fixo, compreendido entre Setembro de 2016 e Maio de 2018. Além disso, os modelos são treinados em modo *batch*, uma única vez e sem atualizações, com período de treino e testes pré-definidos.
- Para as avaliações financeiras os custos transacionais, de operação e incidência de impostos são desconsiderados, com valores representando o rendimento bruto obtido pelas estratégias e algoritmos. A avaliação de custos se mostrou inviável por ser relativa às características do investidor, o meio negociado e valor investido.

5 Análise Experimental

Neste capítulo, retomamos as questões inicialmente propostas:

1. Para o mercado de ações brasileiro, é possível evitar perdas com estratégias automáticas baseadas em mineração de texto de notícias em português?
2. É possível obter dos classificadores explanações de modelo ou instâncias que descrevam os eventos por trás das depreciações?

A primeira questão é tratada através dos experimentos na seção 5.1. Inicialmente são testados e discutidos vários pontos da modelagem como um problema de classificação de textos, avaliando a escolha do método de rotulação, a viabilidade e comparativo de técnicas de seleção de atributos, avaliação de impacto no uso de procedimentos como retornos anormais e extração de radical e um estudo sobre a janela de notícias mais significativa para as predições.

Ao final da seção 5.1 o experimento final, mais abrangente, com todos os classificadores e as melhores parametrizações encontradas, têm seus resultados apresentandos e discutidos, discorrendo sobre os aspectos da primeira pergunta com base nos resultados obtidos. Como o enfoque do trabalho é na observação e análise de quedas de preço, as medidas quantitativas são sempre relativas a previsões de quedas apenas.

A segunda questão é abordada na seção 5.2. As análises buscam a aplicação de técnicas para levantar explanações através dos modelos de classificação, usando interpretação direta e indireta, e avaliando explanações de modelo e de instância. Ao final do capítulo os resultados são debatidos e a pergunta é respondida através dos indícios recuperados dos testes realizados.

5.1 Treinamento de classificadores

Os experimentos foram executados sequencialmente. O experimento 1 (Seção 5.1.1) investiga qual abordagem é mais eficaz na rotulação dos exemplos, discorrendo sobre a análise multiclasse, classificação binária e como sistema de diagnóstico. O experimento discorre também sobre a escolha do limiar de separação entre quedas e baixas no preço, empregado na conversão dos dados numéricos para rótulos.

Como o treinamento dos modelos é demorado, o experimento 2 (Seção 5.1.2) avalia o percentual viável para redução de dimensionalidade no espaço de atributos. É aplicada a seleção de atributos por filtro, usando como métrica o coeficiente de correlação de

Pearson entre a frequência *TF-IDF* das palavras e o preço de cada ação alvo. O objetivo do experimento é determinar um subespaço de atributos que mantenha o desempenho dos modelos reduzindo seu tempo de treinamento.

Dados os resultados em 5.1.2, o experimento 3 (Seção 5.1.3) faz um comparativo entre medidas de redução de dimensionalidade baseadas em filtro, onde o coeficiente de correlação de Pearson tem seus resultados confrontados com as métricas Chi2 e F-Medida, com a finalidade de averiguar qual das métricas conduz a melhores resultados.

O experimento 4 (Seção 5.1.4) traça um comparativo entre o uso de retornos multiplicativos convencionais e retornos multiplicativos anormais. O objetivo é estudar qual das duas abordagens produz os melhores classificadores, considerando tanto métricas de desempenho quanto financeiras. Debate-se também a viabilidade da aplicação dos retornos anormais para o problema proposto.

O experimento 5 (Seção 5.1.5) reduz ainda mais o escopo das buscas avaliando o impacto da janela de notícias no desempenho dos classificadores. Avaliam-se o conjunto de notícias de 1, 5 e 21 dias atrás aplicadas no treinamento dos classificadores, com diferentes abordagens de ponderação.

Com o problema de classificação melhor definido a partir dos testes anteriores, o experimento 6 (Seção 5.1.6) avalia o uso de um conjunto maior de técnicas na resolução do problema, empregando a otimização de hiper-parâmetros através de busca em *grid* e comparando os resultados entre si e também com estratégias de operação tradicionais.

A Tabela 10 resume os parâmetros pré-determinados e os investigados em cada experimento. A primeira coluna define o experimento, a segunda os algoritmos empregados. As colunas "Publicações" e "Retornos" indicam respectivamente as janelas de notícias usadas para treinar os modelos e os prazos de retornos a serem previstos.

Cada combinação de algoritmo, publicação e retorno é empregada na geração de um modelo diferente. Para o experimento 1, por exemplo, a combinação da técnica NB-G, treinada com publicações dos últimos 5 dias, aplicado na predição da direção de retornos para 0 dias representa um dos $(3 * 3 * 3 = 27)$ modelos calibrados.

As colunas "Rotulação" e "*Threshold*" definem o método pelo qual a série de preços é convertida para rótulos, sendo que a primeira determina o domínio de rótulos e a segunda os limiares que separam cada um deles.

Os campos "Subespaço de atributos" e "Filtro" abordam a seleção de atributos, mostrando respectivamente os subespaços de atributos testados. A coluna "Outros" apresenta de maneira genérica demais técnicas empregadas no teste, como retornos anormais e extração de radical.

A última coluna, "*Grid Search*", indica se os algoritmos tiveram seus parâmetros

Tabela 10 – Parâmetros usados em cada experimento, sendo os em cinza alvo de análise e os demais pré-definidos. Parâmetros não usados são marcados com "–".

Exp.	Algoritmos	Publicações	Retornos	Rotulação	*Threshold*	Subespaço de atributos	Filtro	Outro	*Grid Search*
1 (5.1.1)	NB-G *SVM* *KNN*	$\{1, 5, 21\}$	$\{0, 5, 21\}$	Binária, Ternária e Magnitude	$[-5\%, 5\%]$	–	–	–	Não
2 (5.1.2)	NB-G, *SVM* e *KNN*	$\{1, 5, 21\}$	$\{0, 5, 21\}$	Binária	0%	$[2.5\%, 100\%]$	Corr.	–	Não
3 (5.1.3)	NB-G *SVM* e *KNN*	$\{1, 5, 21\}$	$\{0, 5, 21\}$	Binária	0%	2.5%	Corr., Chi2 e F1	–	Não
4 (5.1.4)	NB-G, *SVM* *KNN*	$\{1, 5, 21\}$	$\{0, 5, 21\}$	Binária	0%	2.5%	F1	Retornos anormais	Não
5 (5.1.5)	NB-G, *SVM* e *KNN*	$\{1, 5, 21\}$	$\{0, 5, 21\}$	Binária	0%	2.5%	F1	Publicações ponderadas	Não
6 (5.1.6)	NB-G, *SVM*, *KNN*, AD, FA e RL	$\{1\}$	$\{0, 5, 21\}$	Binária	0%	2.5%	F1	–	Sim

otimizados por busca em *grid* ou se simplesmente assumiram valores padrão da biblioteca Scikit Learn. As células marcadas em cinza na tabela apresentam as configurações testadas em cada experimento, sendo os seguintes executados com as configurações mais viáveis encontradas.

5.1.1 Experimento 1: Definição de rótulos

Para tratar a previsão de preço como um problema de classificação, antes é necessário converter as variações numéricas contínuas para literais nominais. Para tanto, três abordagens foram avaliadas:

1. Classes binárias com valores 0 e 1, sendo consideradas quedas (rótulo 0) retornos abaixo de -1% e altas (rótulo 1) valores maiores ou iguais a este limiar.

2. Classes ternárias $\{-1, 0, 1\}$, onde variações negativas abaixo de -1% são consideradas quedas (rótulo -1), positivas acima de 1% altas (rótulo 1), e valores no intervalo entre -1% e 1% são classificados como estáveis (rótulo 0).

3. Classes de magnitude $\{-2, -1, 0, 1, 2\}$, onde variações entre -2% e +2% são consideradas estáveis (rótulo 0), variações entre +-2% e +-5% são marcadas respectivamente

como altas (rótulo 1) e quedas (rótulo -1) moderadas, e valores acima de +-5% são consideradas altas (rótulo 2) e quedas (rótulo -2) acentuadas.

A terceira abordagem apresentou piora dos resultados sob as métricas avaliadas em função do desbalanceamento de exemplos. Os rótulos definindo quedas e altas acentuadas para o prazo de retorno de 0 dias foram os mais afetados, pois possuem muito menos amostras em relação aos outros (Figura 17). Ademais, o tratamento do problema ganhou complexidade extra nas avaliações devido à análise multiclasse.

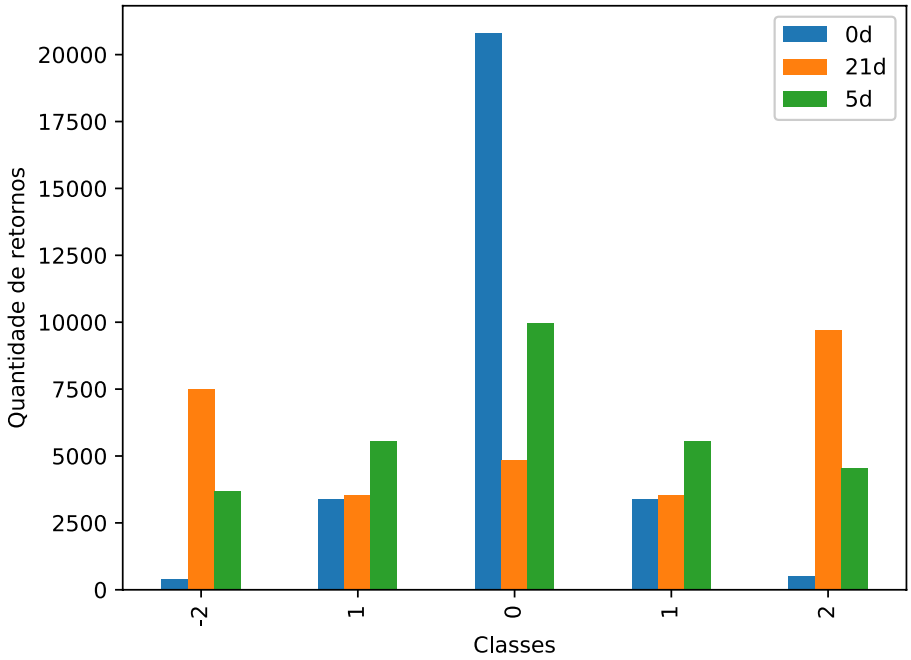


Figura 17 – Quantidade de retornos (exemplos) para cada classe da rotulação por magnitude, considerando período de treino e de teste, com retornos de 0, 5 e 21 dias. Valores somados entre todas as ações.

Fonte: Produzido pelo autor

Os resultados para a primeira e segunda abordagem foram parecidos. Sob uma investigação de quedas somente, os rótulos binários e ternários são equivalentes quando o limiar de separação de baixas é o mesmo, neste caso -1%. No entanto, a rotulação binária invertida, com 1 indicando quedas e 0 altas, aproxima o problema a um sistema de diagósticos, tornando os modelos e análises mais simples.

Foram realizados em seguida testes preliminares avaliando valores de *threshold* do conjunto $\{-1\%, -0.5\%, 0\%, 0.5\%, 1\%\}$, analisados individualmente para as janelas de retorno $\{0, 5, 21\}$, com a técnica NB-G. Os valores médios das medidas de desempenho, entre todas ações, são exibidos na Tabela 11. De maneira geral, há melhora nas medidas de acerto aumentando o limiar entre -1% e 1%, contudo o retorno financeiro apresenta resultados superiores com o valor intermediário 0%.

Tabela 11 – Avaliação de desempenho para classificadores NB-G com rotulação como problema de diagnóstico, com diferentes limiares de separação, para janelas de retorno de 0, 5 e 21 dias. As métricas foram averiguadas para predições relativas aos dados de teste.

Prazo de Retorno	***Threshold***	**Precisão**	**Revocação**	**F-Medida**	**Fator de Retorno Monetário**	**Fator de Perda Monetária**
0d	-0.010	0.4529	0.5093	0.4401	1.2690	0.4053
0d	-0.005	0.6313	0.6126	0.5950	1.4612	0.3226
0d	0.0	0.7147	0.6670	0.6617	1.5454	0.2734
0d	0.005	0.7837	0.7598	0.7513	1.4780	0.2079
0d	0.010	0.8268	0.7561	0.7700	1.2930	0.2170
5d	-0.010	0.7146	0.7510	0.7142	1.2229	0.2036
5d	-0.005	0.7741	0.7855	0.7680	1.2402	0.1781
5d	0.0	0.8037	0.8055	0.7948	1.2433	0.1570
5d	0.005	0.8213	0.8314	0.8175	1.2352	0.1225
5d	0.010	0.8413	0.8301	0.8284	1.2283	0.1073
21d	-0.010	0.8417	0.7580	0.7835	1.1104	0.1385
21d	-0.005	0.8579	0.7657	0.7948	1.1095	0.1272
21d	0.0	0.8758	0.7726	0.8076	1.1114	0.1188
21d	0.005	0.8842	0.7792	0.8140	1.1106	0.1128
21d	0.010	0.8959	0.7762	0.8176	1.1114	0.1032

A definição do *threshold* influencia no balanceamento entre as classes. A tendência de valorização do mercado leva a mais exemplos de alta que de baixa, o alvo da predição, conforme discutido em 4.2.2. Ao ajustar o *threshold* em um valor mais alto, no entanto, instâncias que seriam rotuladas como altas transformam-se em quedas, reajustando o balanceamento e melhorando a precisão e revocação dos modelos. Por outro lado, a definição do limiar entre uma queda e uma alta também é fundamental para a estratégia de operação, pois leva a decisão de compra ou venda.

Ponderadas as avaliações, a escolhida para os testes a seguir é classificação binária, como um sistema de diagóstico de quedas, com limiar de separação entre classes estabelecendo quedas (rótulo 1) para retornos abaixo de 0%, e altas (rótulo 0) quando o retorno for maior ou igual a este limiar.

5.1.2 Experimento 2: Seleção de atributos via correlação

Para o tema proposto, a extensão do texto e a variedade de pontos de vista expressados nos documentos se mostram desafiadores para análise, pois uma mesma notícia pode ter impacto positivo para uma ação, mas negativo para outra. Ademais, o escopo proposto para a análise é multi-documento, analisando as notícias diárias como um conjunto e não individualmente. Assim, a quantidade de termos e pontos de vistas expressados é ainda maior.

Dado um mesmo corpo de palavras diárias, busca-se através deste experimento definir automaticamente qual o subconjunto de palavras mais relevante a cada ação, usando

para isso o atributo alvo e seu coeficiente de correlação de Pearson com cada *feature*. Em resumo, o *corpus* é investigado sobre o ponto de vista da ação avaliada, selecionando as palavras mais relevantes.

Este teste foi executado com os classificadores Naive Bayes Gaussiano (NB-G), K Vizinhos Mais Próximos (*KNN*) e Máquinas de Vetores de Suporte (*SVM*). Os algoritmos foram aplicados com seus parâmetros padrões da biblioteca Scikit Learn.

O *dataset* escolhido foi a representação *TF-IDF*, com a seleção de atributos feita através da análise da correlação entre a frequência *TF-IDF* e as variações diárias de preço. Foram selecionadas as seguintes porcentagens do espaço original de atributos: 2.5%, 5%, 10%, 35%, 75% e 100%. Para cada porcentagem, os classificadores foram testados contra as janelas de retorno de 0, 5 e 21 dias úteis, contrapostos às publicações de 1, 5 e 21 dias atrás.

As médias das medidas avaliadas, entre todas as ações e janelas temporais, são dadas na Tabela 12, onde as porcentagens nas colunas retratam o percentual de atributos mantidos, após a aplicação da seleção por correlação.

Tabela 12 – Métricas de avaliação para diferentes subgrupos de atributos

Classificador	Métrica	2.5%	5%	10%	35%	75%	100%	Desv. Padrão
KNN	F. Perda Financeira	0.3194	0.3218	0.3294	0.3577	0.3876	0.3868	0.0315
KNN	F. Retorno Financeiro	0.9901	0.9842	0.9688	0.9620	0.9483	0.9400	0.0196
KNN	Precisão	0.5479	0.5367	0.5178	0.4738	0.4385	0.4021	0.0581
KNN	Revocação	0.4357	0.4447	0.4376	0.4027	0.3796	0.3934	0.0272
KNN	F-Medida	0.4137	0.4090	0.3973	0.3645	0.3391	0.3377	0.0344
NB-G	F. Perda Financeira	0.2074	0.2104	0.2122	0.2258	0.3242	0.3796	0.0736
NB-G	F. Retorno Financeiro	1.2234	1.2519	1.2747	1.2422	1.0544	0.9488	0.1326
NB-G	Precisão	0.6950	0.7174	0.7413	0.7344	0.6390	0.5362	0.0783
NB-G	Revocação	0.6102	0.6251	0.6446	0.6229	0.4507	0.3605	0.1177
NB-G	F-Medida	0.6235	0.6414	0.6603	0.6347	0.4499	0.3443	0.1303
SVM	F. Perda Financeira	0.4152	0.4307	0.4443	0.4754	0.4930	0.4945	0.0335
SVM	F. Retorno Financeiro	1.0362	1.0303	1.0210	0.9878	0.9518	0.9399	0.0414
SVM	Precisão	0.5491	0.5426	0.5368	0.4341	0.3106	0.2750	0.1231
SVM	Revocação	0.3259	0.3059	0.2928	0.2580	0.2354	0.2304	0.0393
SVM	F-Medida	0.3135	0.2968	0.2841	0.2408	0.2063	0.1969	0.0489

A Tabela 13 mostra a quantidade de termos mantidos no espaço de atributo após a aplicação do filtro. O tamanho do espaço de atributos é variável dependendo da janela de publicação considerada, isto se deve ao fato de que, para as diferentes janelas consideradas, distintos valores de *TF-IDF* são produzidos. Em geral, considerando apenas publicações de D_{-1} mais termos têm a frequência constante, logo mais termos são filtrados pela pré seleção por baixa variância.

Na Tabela 12 é possível identificar que, em reduções para até 10% dos atributos, não só se mantém o desempenho dos classificadores como também, na maioria dos casos, há ganhos significativos com o filtro. Abaixo de 10% o que se observa são oscilações mais modestas. Para analisar os dados da Tabela 12 sob formalidade estatística foi aplicado

Tabela 13 – Tamanho do vocabulário (quantidade de *features*) nos subespaços de atributos.

Publicações	**2.5%**	**5%**	**10%**	**35%**	**75%**	**100%**
D_{-1}	7106	14211	28423	99482	213177	284237
D_{-5}	8190	16380	32760	114660	245700	327600
D_{-21}	8190	16380	32760	114660	245700	327600

o teste de Friedman, com o objetivo de determinar se há diferença significativa entre os resultados, para os diferentes subespaços de atributos.

O teste foi executado para cada uma das métricas individualmente, sendo que para uma significância (*alpha*) de 5%, a hipótese nula foi rejeitada entre todos os subgrupos de atributos, logo ao menos um dos sub-conjuntos de *features* apresenta valores estatisticamente diferentes dos demais. No entanto, ao reduzir a comparação aos dados de 2.5%, 5% e 10% somente, o *p-value* sobe a 0.049, muito próximo ao alpha de 5%. Os dados são exibidos na Tabela 14.

Tabela 14 – Valor de *p-value*, segundo teste de Friedman, para as diferentes métricas comparadas nos experimentos, com diferentes subconjuntos dos atributos originais.

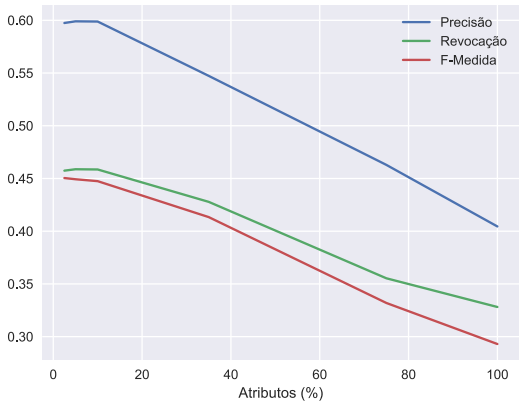
Grupo	**F. Perda F.**	**F. Retorno F.**	**Precisão**	**Revocação**	**F-Medida**
{2.5%,5%,10%,35%,75%,100%}	0.0024	0.0057	0.0111	0.0057	0.0024
{2.5%,5%,10%}	0.0497	0.0497	0.0497	.0497	0.0497

Os resultados podem ser verificados também nas figuras 18 e 19, onde são apresentados a F-Medida, revocação, precisão, perda realizada e retorno monetário médio entre os classificadores. Em termos computacionais e de complexidade do problema, ao considerar 10% dos atributos originais com resultados positivos (redução portanto de 90% no vocabulário original), sinaliza-se a possibilidade de simplificar os testes.

Os resultados podem ser justificados pelo caráter genérico dos documentos, pois no *dataset* original as palavras são fornecidas ao classificador sem análise prévia, sem julgar a pertinência da notícia. Ao aplicar a correlação com o preço, por outro lado, são descartados termos menos relevantes, tornando o classificador mais simples. Além disso, a redução da quantidade de características reduz o super ajustamento dos classificadores aos dados de treino, promovendo melhorias nos testes.

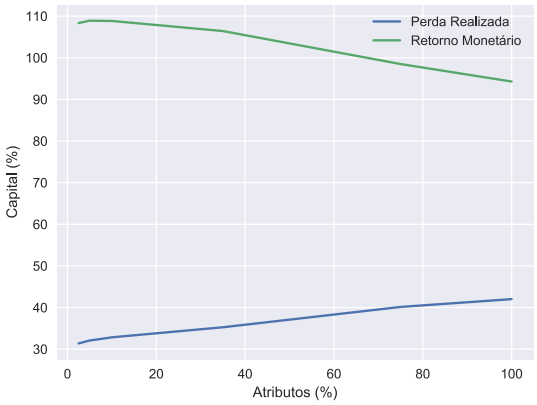
Para o próximo experimento, seção 5.1.3, o subgrupo de 2.5% do vocabulário foi escolhido para nova avaliação, desta vez com mais métricas aplicadas no filtro. Essa porcentagem será adotada por apresentar bons resultados e reduzir o tempo de execução

Figura 18 – Medidas de desempenho para cada percentual dos atributos originais. Média entre *KNN* e NB-G



Fonte: Produzido pelo autor

Figura 19 – Medidas monetárias para cada percentual dos atributos originais. Média entre *KNN* e NB-G



Fonte: Produzido pelo autor

das avaliações.

5.1.3 Experimento 3: Comparativo de métricas de filtro

Neste teste a métrica de correlação de Pearson foi comparada às medidas de seleção Chi2 e F-Medida. O coeficiente de correlação foi aplicado em relação a série de preço, enquanto as métricas F-Medida e Chi2 foram usadas sobre o atributo alvo (queda/alta). Para a métrica Chi2, os valores de *TF-IDF* foram discretizados para computar a importância dos atributos e efetuar a seleção, após a redução os algoritmos foram treinados e testados sobre os valores originais.

Dados os resultados em 5.1.2, foi usado o percentual de 2.5% dos atributos, aplicado aos mesmos algoritmos de classificação Naïve Bayes Gaussiano (NB-G), K Vizinhos Mais Próximos (*KNN*) e Máquinas de Vetores de Suporte (*SVM*), com as janelas $\{1, 5, 21\}$ para publicações e $\{0, 5, 21\}$ para retornos.

As medidas de avaliação são mostradas na Tabela 15. De maneira geral, exceto para o *KNN*, o coeficiente de correlação foi a métrica que apresentou os piores resultados, com as medidas F-Medida e Chi2 levando a um melhor desempenho.

A análise estatística dos resultados foi feita por meio de *T-Test*, analisado as métricas de seleção par a par, avaliadas para as medidas de desempenho F1 e ganho monetário. O *p-value* obtido em cada teste é exibido na Tabela 16.

Através dos valores de *p-value* mostrados na Tabela 16 é possível perceber que a seleção por F-Medida e Chi2 obtiveram resultados mais próximos, enquanto o coeficiente de correlação teve resultados mais distintos.

Tabela 15 – Desempenho dos classificadores com 2.5% de seleção de atributos, por diferentes métricas de filtro.

Classificador	Seleção	Precisão	Revocação	F-Medida	F. Retorno Monetário	F. Perda Monetária
NB-G	Correlação	0.6950	0.6102	0.6235	1.2234	0.2074
NB-G	Chi2	0.7981	0.7484	0.7547	1.3001	0.1831
NB-G	F-Medida	0.7961	0.7459	0.7528	1.2925	0.1767
KNN	Correlação	0.5479	0.4357	0.4137	0.9901	0.3194
KNN	Chi2	0.5483	0.4071	0.3820	0.9782	0.3659
KNN	F-Medida	0.5432	0.4240	0.4050	0.9843	0.3453
SVM	Correlação	0.5491	0.3259	0.3135	1.0362	0.4152
SVM	Chi2	0.7630	0.4623	0.5021	1.1677	0.3463
SVM	F-Medida	0.7473	0.4573	0.5002	1.1720	0.3503

Tabela 16 – Valor de *p-value* avaliado por meio de *T-Test*, par a par, paras as medidas de desempenho F1 e Ganho Monetário.

F1			
	Chi2	**F-Medida**	**Correlação**
Chi2	1.0000	0.9682	0.5382
F-Medida	0.9682	1.0000	0.4998
Correlação	0.5382	0.4998	1.0000
Fator de ganho monetário			
	Chi2	**F-Medida**	**Correlação**
Chi2	1.0000	0.9945	0.6073
F-Medida	0.9945	1.0000	0.5935
Correlação	0.6073	0.5935	1.0000

Para os experimentos seguintes será adotado o filtro por F-Medida, que proporciona bons resultados, custo computacional reduzido e não requer a discretização das frequências *TF-IDF*, como no caso do Chi2.

5.1.4 Experimento 4: Retornos anormais

Trabalhos como (64) e (57) citam o uso de retornos anormais como uma maneira de descartar movimentos do mercado como um todo, focando na observação de padrões no comportamente dos preços das ações isoladamente.

Com o objetivo de comparar o desempenho dos retornos anormais em relação aos convencionais, foi conduzido um experimento com os algoritmos Naïve Bayes Gaussiano (NB-G), K Vizinhos Mais Próximos (*KNN*) e Máquinas de Vetores de Suporte (*SVM*), com as Janelas $\{1, 5, 21\}$ para publicações e $\{0, 5, 21\}$ para retornos. Foi usado no experimento o dataset *TF-IDF* submetido à seleção de 2.5% dos atributos por filtro via F-Medida.

Foram aplicados retornos anormais tendo como base a série histórica do índice Ibovespa. Como os retornos anormais descrevem oscilações em relação ao índice base, seu emprego no cálculo das métricas de avaliação financeira é inviável, logo este foi usado apenas no treino dos classificadores, com as métricas financeiras computadas sobre retornos

convencionais .

Na Tabela 17 são apresentados os resultados em relação ao mesmo teste executado com metodologia convencional de retornos. O treino dos modelos usando retornos anormais apresentou melhorias em quase todas as métricas, mas o retorno monetário para os classificadores *SVM* e NB-G caiu.

Tabela 17 – Comparativo de resultados entre o uso de retornos convencionais e anormais durante o treino de classificadores.

Classif.	**Tipo de Retorno**	**Precisão**	**Revocação**	**F-Medida**	**F. Retorno Monetário**	**F. Perda Monetário**
KNN	Anormal	0.5764	0.5865	0.5202	0.9916	0.2454
KNN	Convencional	0.5483	0.4071	0.3820	0.9782	0.3659
NB-G	Anormal	0.7876	0.7926	0.7746	1.2259	0.1610
NB-G	Convencional	0.7981	0.7484	0.7547	1.3001	0.1831
SVM	Anormal	0.7096	0.7897	0.7021	1.1173	0.1468
SVM	Convencional	0.7630	0.4623	0.5021	1.1677	0.3463

Após análise minunciosa, concluí-se que os retornos anormais potencializam a recuperação de padrões mais refinados, o que se confirma pela melhora nas métricas de acurácia. Por outro lado, estes levam a identificação de padrões em relação à série do índice referência, o Ibovespa, e não necessariamente a variação de preço em si.

Olhando predições caso a caso pode-se observar situações em que quedas em relação ao índice base são antecipadas corretamente, entretanto a posição vendida é assumida de maneira precipitada, pois apesar do decréscimo em relação ao retorno do índice, a variação de preço ocorrida ainda foi positiva. Este comportamento é ilustrado no exemplo da Tabela 18, com predições de 0 dias para o papel PETR4.SA.

Na Tabela 18 as ordens “C” indicam posição comprada e “V” vendida. Para o dia 18/09/2017 nota-se que o retorno anormal do papel, em relação ao índice, foi negativo, contudo o retorno normal ainda mostrou valorização. Treinados com os retornos anormais todos os classificadores fizeram corretamente a previsão de venda, entretanto a ordem esperada, mais eficaz para a estratégia, era manter a posição comprada. O caso inverso é dado no dia 28/09/2017, onde NB-G fez a predição correta, porém com resultado negativo.

Os casos de oportunidades de ganho perdidas com os retornos anormais se intensificam quanto melhor a qualidade do classificador, pois altas que seriam previstas corretamente com retornos normais são previstas como quedas em relação ao índice de referência. Isso explica o fato do comportamento ser observado apenas no *SVM* e NB-G, pois *KNN* sequer consegue desempenho significativo com retornos comuns.

Apesar dos retornos anormais levar a melhores métricas quantitativas, esta me-

Tabela 18 – Análise de predições de 0 dias para PERT4.SA. Linhas destacadas mostram predições corretas em relação ao índice de referência, mas erradas para estratégia de operação.

Data	Retorno Anormal	Retorno Normal	Ordem Esperada	Ordem NB-G	Ordem *KNN*	Ordem *SVC*
2017-09-15	-0.0167	-0.0020	V	V	V	V
2017-09-18	-0.0017	0.0013	C	V	V	V
2017-09-19	0.0029	0.0026	C	C	V	C
2017-09-20	0.0403	0.0407	C	C	V	C
2017-09-21	-0.0060	-0.0114	V	V	V	V
2017-09-22	0.0088	0.0058	C	C	V	C
2017-09-25	0.0156	0.0032	C	C	V	C
2017-09-26	-0.0172	-0.0189	V	V	V	V
2017-09-27	-0.0173	-0.0248	V	V	V	V
2017-09-28	0.0013	-0.0020	V	C	V	V
2017-09-29	-0.0227	-0.0129	V	V	V	V
2017-10-02	0.0129	0.0138	C	C	C	V
2017-10-03	-0.0079	0.0244	C	V	C	V
2017-10-04	-0.0129	-0.0151	V	V	V	V
2017-10-05	0.0009	0.0013	C	C	V	C
2017-10-06	0.0093	0.0019	C	C	V	C
2017-10-09	0.0222	0.0179	C	C	C	V
2017-10-10	-0.0115	0.0037	C	V	V	V
2017-10-11	-0.0023	-0.0056	V	V	V	V

todologia se mostra inviável se aplicada na modelagem de classificadores para uso como estratégia de operação. Por este motivo será descartado para os testes seguintes.

5.1.5 Experimento 5: Janela de publicações

Este experimento busca determinar qual o prazo de notícias mais relevante na tomada de decisão dos investidores, avaliando publicações do dia anterior à análise de retornos, dos 5 dias anteriores e também para 21 dias anteriores.

Quanto ao tratamento, foram testados pesos uniformes entre todas as notícias e ponderação decrescente conforme a data de publicação. Na abordagem ponderada, os pesos foram testados tanto aplicados na frequência *TF*, antes do cálculo do *TF-IDF*, quanto posteriormente, na soma ponderada do *TF-IDF* para cada uma das datas. Em resumo, as representações de dados para o experimento foram organizadas conforme a Tabela 19.

Os testes foram feitos com os algoritmos Naïve Bayes Gaussiano (NB-G), K Vizinhos Mais Próximos (*KNN*) e Máquinas de Vetores de Suporte (*SVM*), com os janelas $\{1, 5, 21\}$. O espaço de atributos foi previamente filtrado para 2.5% do tamanho original, pela métrica F-Medida. Os resultados são exibidos na Figura 20.

Pelos dados na Figura 20 é possível notar que para prazos curtos, de 0 e 5 dias, as publicações do anterior são as mais impactantes na movimentação dos preços, pois a representação TD-IDF-1 levou aos melhores resultados. Para o prazo de 21 dias, contudo, há a indicação de que um grupo maior de publicações passadas conduz a melhores resultados, com melhorias para as representações TF-IDF-5 e TF-IDF-21 ante a TF-IDF-1.

Tabela 19 – Abordagens para representação de notícias.

Identificador	**Prazo**	**Fórmula**	**Descrição**
TF-IDF-1	1	$Freq(t_i) = tfidf = log(tf(p_i, d) + 1) * log(\frac{\|D\|+1}{df(p_i)+1})$	Fórmula original de *TF-IDF*, conforme (1).
TF-IDF-5	5	$Freq(t_i) = \sum_{j=0}^{4} tfidf(t_{i-j})$	*TF-IDF* com notícias replicadas por 5 dias.
TF-IDF-21	21	$Freq(t_i) = \sum_{j=0}^{20} tfidf(t_{i-j})$	*TF-IDF* com notícias replicadas por 21 dias.
EW-TF-IDF-5	5	$Freq(t_i) = \sum_{j=0}^{4} tfidf(t_{i-j}) * w(i-j)$	*TF-IDF* com notícias replicadas por 5 dias e peso exponencial decrescente.
EW-TF-IDF-21	21	$Freq(t_i) = \sum_{j=0}^{20} tfidf(t_{i-j}) * w(i-j)$	*TF-IDF* com notícias replicadas por 21 dias e peso exponencial decrescente.
EWTF-IDF-5	5	$Freq(t_i) = \sum_{j=0}^{4} log(tf(p_i, t_i) * w(i-j) + 1) * log(\frac{\|D\|+1}{df(p_i)+1})$	*TF-IDF* com notícias replicadas por 5 dias e peso exponencial decrescente aplicado à frequência *TF*.
EWTF-IDF-21	21	$Freq(t_i) = \sum_{j=0}^{20} log(tf(p_i, t_i) * w(i-j) + 1) * log(\frac{\|D\|+1}{df(p_i)+1})$	*TF-IDF* com notícias replicadas por 21 dias e peso exponencial decrescente aplicado à frequência *TF*.

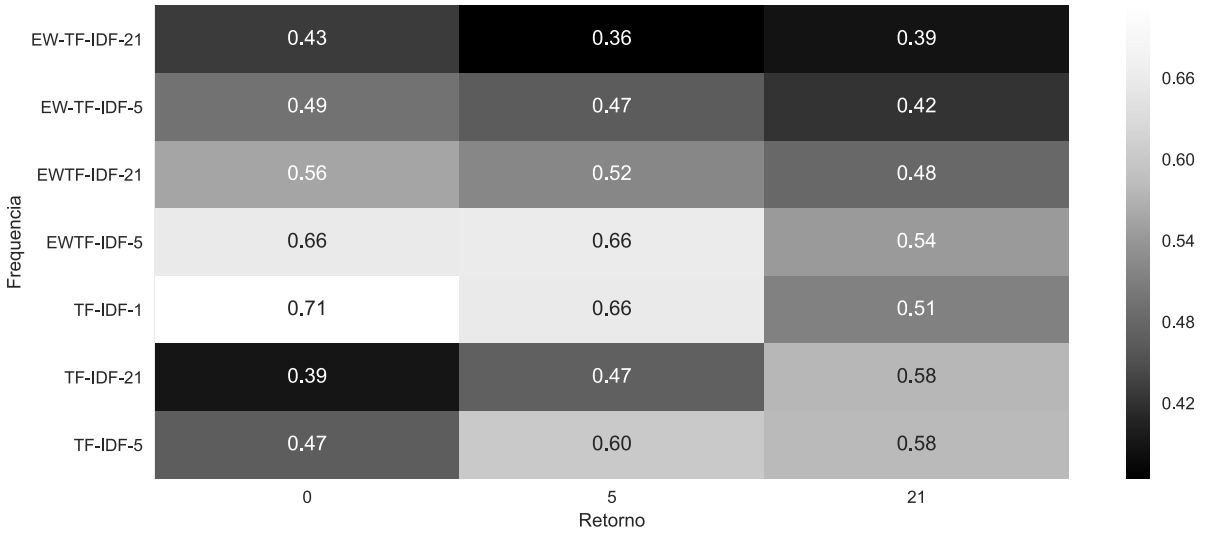


Figura 20 – F-Medida médio entre os classificadores treinados sob diferentes representações, com e sem replicação de notícias. Tons claros indicam maior desempenho, escuros pior.

Fonte: Produzido pelo autor

É possível observar também que as tentativas de atribuir pesos maiores às notícias mais recentes, seja na frequência *TF* ou *TF-IDF* diretamente, não resultam em melhoria, o que fica claro na terceira coluna da tabela. Dados os resultados insatisfatórios, os testes seguintes são executados usando apenas a representação TF-IDF-1.

5.1.6 Experimento 6: Algoritmos e busca em *grid*

Dadas as configurações com melhores resultados obtidas nos testes anteriores, este experimento busca avaliar o desempenho de diferentes algoritmos na tarefa de classificação. A lista dos algoritmos aplicados é dada na Tabela 20.

Tabela 20 – Algoritmos e parâmetros para busca em *grid*.

Algoritmo	**Abreviação**	**Parâmetros**
Naïve Bayes Gaussiano	NB-G	Padrão
Árvores de Decisão	AD	Critério={Gini, Entropia}; Balanceamento={Automático, Nenhum}; Exemplos Min. em Folhas={1, 3, 5}
Floresta Aleatória	FA	Critério={Gini, Entropia}; Estimadores={250}; Balanceamento={Automático, Nenhum}; Exemplos Min. em Folhas={1, 3, 5}
Multilayer Perceptron	MLP	*Shuffle*={Sim, Não}; Taxa de Aprendizado={Adaptativa}; Alpha={0.0001, 0.01, 1, 10}; Épocas={100}; Neuronios Cam. Oculta={100, 50}
Máquinas de Vetores de Suporte	SVM	Kernel={RBF}; C={0.001,0.01,0.1,1}; Gamma={0.001,0.01,0.1,1,Auto}; Balanceamento={Automático, Nenhum}
Regressão Logística	RL	Solver={lbfgs}; C={0.001,0.01,0.1,1,10}
K Vizinhos mais Próximos	KNN	K={3,8,15}; Pesos={Uniforme, Distância}

Os testes foram executados conforme os protocolos anteriores, com redução de dimensionalidade a 2.5% por filtro em F-Medida, classificação binária com limiar em 0% e retornos {0, 5, 21}. Os algoritmos foram treinados com busca em *grid* de parâmetros, novamente com um modelo por ação e janela de retorno. Os parâmetros testados são exibidos na coluna "Parâmetros" da Tabela 20.

Dados os resultados da Seção 5.1.5, foram usadas apenas publicações do dia anterior a análise (D_{-1}) para o treinamento dos modelos. No Apêndice A são apresentadas as distribuições dos dados entre as classes, para treino e teste, com os janelas de retorno de D_0, D_{+5} e D_{+21}.

A busca em *grid* foi efetuada três vezes para otimização de diferentes métricas: ganho monetário percentual, perda monetária percentual e F-Medida. As métricas foram usadas como medidas para maximização, sendo que no caso da perda foi empregado o inverso: $1/perda$.

Os resultados para cada classificador, e para cada medida usada na otimização, são apresentados na Tabela 21. As métricas na tabela indicam a média entre todas as ações e todos os intervalos de tempo.

Tabela 21 – Resultados obtidos para todos os classificadores, com parâmetros otimizados por busca em *grid*.

Classificador	Otimização	Precisão	Revocação	F-Medida	F. Retorno Monetária	F. Perda Monetária
AD	F-Medida	0.5497	0.4315	0.4613	0.9954	0.2118
AD	Ganho Mon. Perc.	0.5479	0.4274	0.4601	1.0033	0.2016
AD	Perda Mon. Perc.	0.5531	0.4336	0.4662	1.0040	0.2030
FA	F-Medida	0.6219	0.3967	0.4405	1.0510	0.2811
FA	Ganho Mon. Perc.	0.6348	0.3776	0.4284	1.0403	0.3057
FA	Perda Mon. Perc.	0.6247	0.3345	0.4010	1.0584	0.3166
KNN	F-Medida	0.5526	0.5023	0.4322	0.9943	0.2594
KNN	Ganho Mon. Perc.	0.4725	0.4376	0.3779	0.9802	0.3080
KNN	Perda Mon. Perc.	0.4480	0.3170	0.2825	0.9453	0.4019
MLP	F-Medida	0.8852	0.7569	0.7877	1.5414	0.1455
MLP	Ganho Mon. Perc.	0.8927	0.7413	0.7774	1.5209	0.1508
MLP	Perda Mon. Perc.	0.8875	0.7439	0.7759	1.5304	0.1501
NB-G	F-Medida	0.9268	0.8326	0.8643	1.5604	0.0840
NB-G	Ganho Mon. Perc.	0.9268	0.8326	0.8643	1.5604	0.0840
NB-G	Perda Mon. Perc.	0.9268	0.8326	0.8643	1.5604	0.0840
RL	F-Medida	0.8802	0.7150	0.7561	1.4814	0.1641
RL	Ganho Mon. Perc.	0.8684	0.6946	0.7407	1.4481	0.1810
RL	Perda Mon. Perc.	0.8711	0.6830	0.7324	1.4402	0.1889
SVM	F-Medida	0.7163	0.6968	0.6696	1.2798	0.1685
SVM	Ganho Mon. Perc.	0.4622	0.5195	0.4509	1.0781	0.2811
SVM	Perda Mon. Perc.	0.3235	0.3743	0.3069	0.9672	0.3971

Nota-se que, apesar de especializadas, a otimização por meio de métricas financeiras não contribui para a melhoria dos resultados do classificador. A busca em *grid* empregando otimização por F-Medida conduziu a melhores resultados para a maioria dos algoritmos, com diferenças expressivas em relação às outras otimizações.

Com relação a perdas monetárias, considerando a otimização por F-Medida apenas, todos os algoritmos obtiveram desempenhos razoáveis, com o melhor deles, NB-G, conseguindo a média de 8.40% entre todas as ações e o pior, *KNN*, 25.94%.

Em termos de ganhos, por outro lado, os algoritmos SVM, NB-G, MLP e RL conseguem prover rendimentos muito mais altos mantendo baixo o limiar de perdas. Observando casos isolados, nota-se que a combinação de precisão e revocação elevadas levam estes algoritmos a antecipar quedas devidamente, sem assumir uma posição muito conservadora.

Os algoritmos FA, AD e KNN apresentaram maior dificuldade em separar as classes, com o último inclusive não conseguindo evitar resultados negativos na maioria dos casos. A Tabela 22 mostra como se comporta o desempenho dos algoritmos nas diferentes fases: treino, validação e teste. Os dados representam valores médios entre as ações e janelas temporais.

Tabela 22 – Evolução do desempenho em F-Medida nas fases de treino, validação e teste dos algoritmos. Valor médio entre as janelas de retorno.

Classificador	**Treino**	**Validação**	**Teste**
AD	0.9508	0.5070	0.4613
NB-G	0.9367	0.8751	0.8643
KNN	0.7880	0.3262	0.4322
RL	0.9632	0.7705	0.7561
MLP	0.9921	0.8772	0.7877
FA	0.9191	0.5127	0.4405
SVM	0.7733	0.6375	0.6696

Observa-se na Tabela 22 que os algoritmos AD, KNN e FA obtiveram ótimos resultados na fase de treino, mas piora significativa em validação e teste, evidenciando super ajustamento aos dados de treino. Os demais algoritmos também apresentam super ajustamento, contudo em menor escala e mantendo desempenhos razoáveis. O algoritmo NB-G se mostrou o menos susceptível ao super ajustamento, com desempenho próximo entre treino e teste.

As figuras 21 e 22 mostram o desempenho financeiro dos classificadores em dois casos isolados, a ação com o pior e a de melhor resultado no período, respectivamente. Ambos os casos foram avaliados para retornos entre a abertura e fechamento do mercado no mesmo dia.

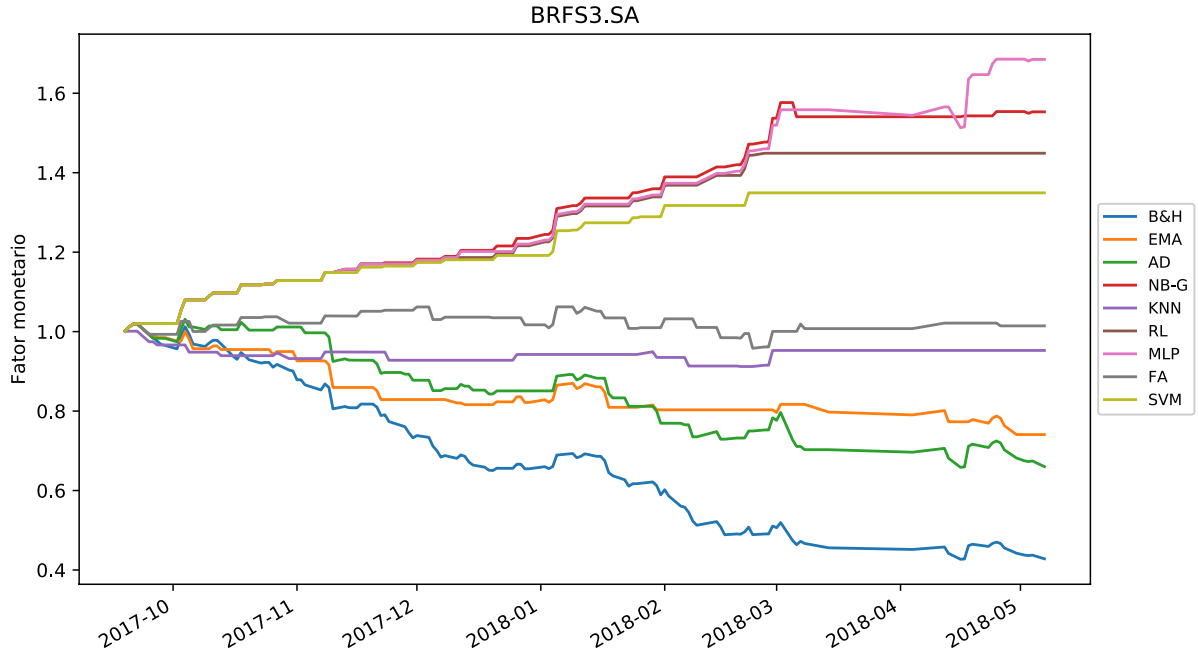


Figura 21 – Classificadores avaliados contra a ação com piores retornos no período: BRFS3.SA.

Fonte: Produzido pelo autor

No caso da ação BRFS3.SA o gráfico na Figura 21 demonstra que todos os algoritmos conseguem ao menos amenizar as perdas ocorridas no período, que pela estratégia *Buy & Hold* totaliza quase 60% do capital investido. Nota-se também que com exceção do algoritmo AD, o qual pode não ter se favorecido pelo tratamento de dados, todas as demais abordagens baseadas em aprendizagem de máquina superaram as duas técnicas tradicionais, *Buy & Hold* e EMA.

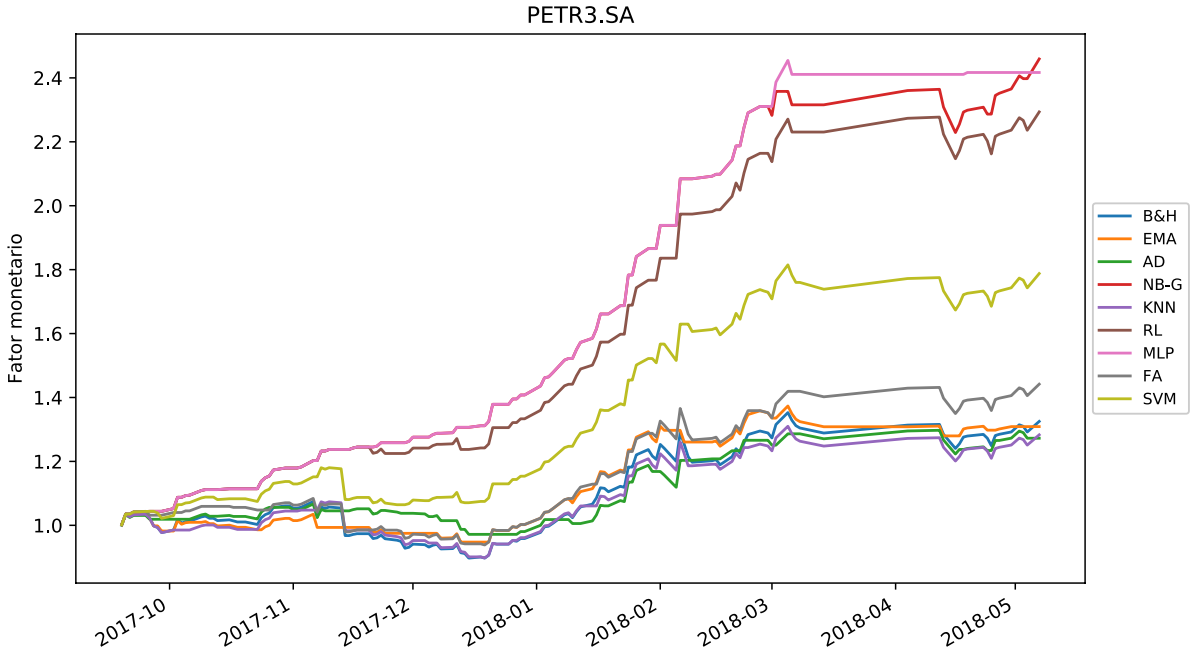


Figura 22 – Classificadores avaliados contra a ação com melhores retornos no período: PETR3.SA.

Fonte: Produzido pelo autor

Três técnicas no entanto se sobressaem, conseguindo não só evitar a maioria das quedas mas também produzindo ganhos com as oscilações positivas de preço, são elas: NB-G, MLP e RL. Observando as figuras 21 e 22, nota-se que estes conseguem adiantar e evitar quedas com precisão e quase instantaneamente na maioria dos casos, diferente da técnica EMA que notoriamente sofre atraso ao antecipar quedas acentuadas. Os dois casos podem ser constatados entre Novembro e Dezembro de 2017, para ambas ações.

A Figura 23 compara o gráfico *boxplot* de cada uma das estratégias. A métrica avaliada é a distribuição dos fatores de retornos, sob análise de publicações de D_{-1} e retornos em D_0, mostrando o valor médio entre as ações. A Figura 24 apresenta a perda percentual, a porcentagem das quedas sofridas entre as ocorridas, analisada sob as mesmas circunstâncias.

Estendendo os dados apresentados para as duas ações nas figuras 21 e 22, o gráfico da Figura 23 mostra que, independente da ação avaliada, os resultados obtidos pelos classificadores NB-G, MLP e RL foram sempre positivos no período testado. Com relação

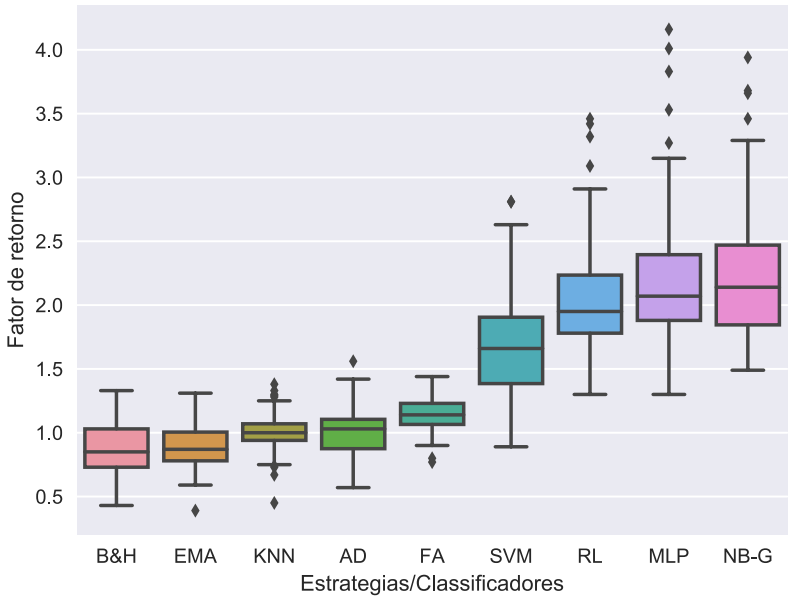


Figura 23 – Distribuição de retornos entre classificadores e estratégias, para publicações do dia anterior e prazo em D_0.

Fonte: Produzido pelo autor

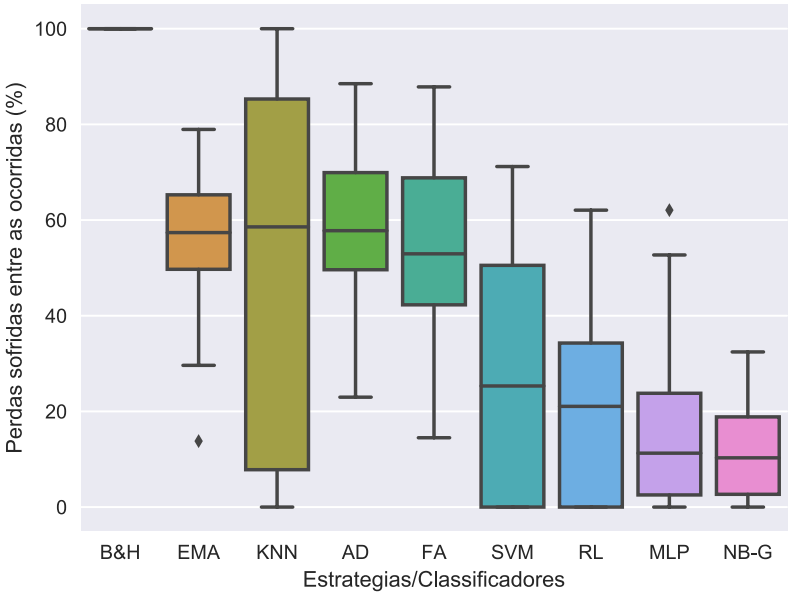


Figura 24 – Perda percentual entre classificadores e estratégias, para publicações do dia anterior e prazo em D_0.

Fonte: Produzido pelo autor

as perdas da Figura 24, as técnicas com melhores retornos são também as que promovem o menor risco, com NB-G sofrendo em média apenas 10% das perdas monetárias do período.

Quanto aos parâmetros de treinamento, não houve consenso entre quais desempenham melhor, mas cada ação e janela temporal apresentou atributos específicos como melhores na busca em *grid*. Isso se justifica pois apesar da mesma abordagem, cada ação e janela temporal representa um sistema diferente a ser previsto, logo é razoável que se otimizem com parâmetros também diferentes.

5.1.7 Discussão dos resultados

Para o teste com melhor desempenho na Seção 5.1.6, usando F-Medida, as figuras 25 e 26 exibem respectivamente a medida de desempenho F-Medida e o ganho financeiro, para diferentes janelas de retornos. Os dados apresentados são médios entre todas as ações e os classificadores NB-G, MLP e RL.

Figura 25 – Métricas financeiras consolidadas para diferentes prazos retornos.

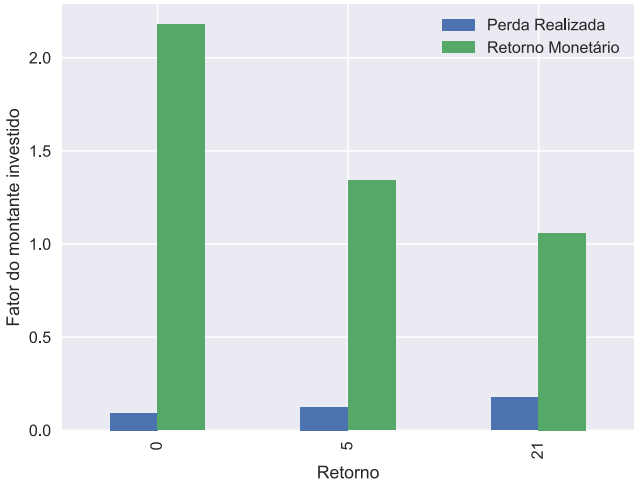
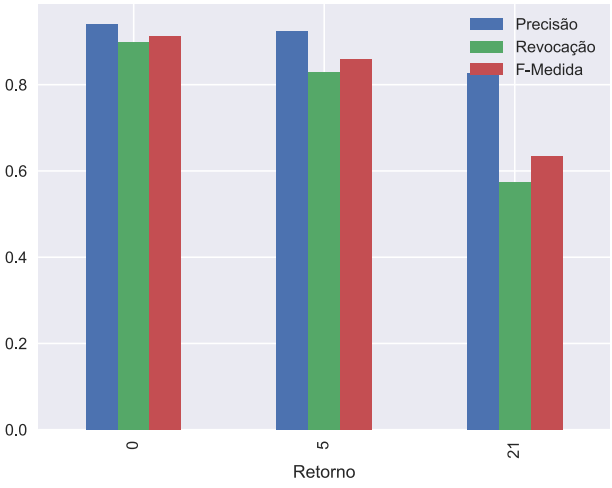


Figura 26 – Métricas quantitativas consolidadas para diferentes prazos retornos.



Fonte: Produzido pelo autor

Para retornos os melhores resultados se concentram no curto prazo, sendo que o ganho financeiro para análises de 21 dias cai consideravelmente. Como estratégia de operação, ao averiguar as predições caso a caso, as análises de longo prazo são prejudicadas pelos extensos períodos sob a posição vendida, onde oportunidades de ganho são desperdiçadas.

Apesar das predições de longo prazo serem menos viáveis para operação, estas podem contribuir para a seleção dos ativos para compor a carteira de investimentos, na otimização do portifólio, evitando papéis com tendência de quedas a longo prazo e reduzindo os riscos.

Os experimentos foram executados sobre todas as ações que compunham o índice IBovespa no mês de dezembro de 2018, no entanto supõe-se de antemão que diferentes empresas, ou ramos de negócios, têm diferentes sensibilidades às notícias. Assim, buscou-se identificar quais foram as ações mais influenciadas, portanto mais promissoras ao encontro de padrões.

A Tabela 23 mostra as médias das principais métricas para cada segmento de negócio. O portal da bolsa de valores B3 fornece a categorização de cada um dos ativos negociados em segmentos de atividade, em níveis macro e micro[1], a qual foi usada para consolidar os resultados da tabela. Os dados se referem a previsões de retornos em D_0 apenas, com valores médios entre os quatro melhores algoritmos, NB-G, SVM, MLP e RL, ordenados pela métrica F-Medida.

Tabela 23 – Medidas de avaliação agrupadas por segmento de negócio. Destaque aos três setores onde menos perdas foram evitadas e máximos e mínimos para perda percentual e retorno monetário.

Segmento	*Precisão*	*Revocação*	*F-Medida*	*Perda (%)*	*F. Retorno M.*	*Buy & Hold*
Papel e Celulose	0.9467	0.5867	0.6833	43.6782	1.3733	1.1200
Minerais Metálicos	0.9050	0.8817	0.8867	15.7757	2.2767	1.1850
Cervejas e Refrigerantes	1.0000	0.8100	0.8900	33.3333	1.4533	1.0100
Produtos Diversos	0.9733	0.8267	0.8917	23.2884	2.8667	0.8600
Serviços Educacionais	0.9533	0.8517	0.8917	24.9610	2.2967	0.7850
Bancos	0.9811	0.8406	0.8989	24.3737	1.9322	0.9883
Exploração, Refino e Distribuição	0.9650	0.8567	0.9033	20.9193	2.0842	1.0900
Análises e Diagnósticos	0.9767	0.8500	0.9067	16.8615	1.9767	0.7300
Telecomunicações	0.9850	0.8450	0.9067	18.8797	1.8550	1.0250
Tecidos, Vestuário e Calçados	0.8733	0.9567	0.9067	5.0847	2.2400	1.0400
Seguradoras	0.8667	0.9633	0.9067	6.8027	1.5633	0.8900
Eletrodomésticos	0.9800	0.8667	0.9167	26.7884	2.9817	0.5400
Aluguel de carros	0.9200	0.9267	0.9167	14.4444	2.2067	1.0400
Siderurgia	0.9358	0.9142	0.9200	12.6016	3.0858	0.7375
Alimentos	0.9133	0.9333	0.9200	5.9140	2.1567	0.9100
Medicamentos e Outros Produtos	0.9700	0.8850	0.9200	19.3621	1.8717	0.9950
Transporte Ferroviário	0.8933	0.9567	0.9200	4.9020	2.9800	1.0800
Serviços Financeiros Diversos	0.9083	0.9433	0.9217	9.5192	1.9883	0.8850
Carnes e Derivados	0.9456	0.9100	0.9244	13.3762	1.9144	0.6367
Programas de Fidelização	0.9933	0.8600	0.9267	15.4930	2.2967	0.7500
Motores, Compressores e Outros	0.9300	0.9333	0.9267	9.2308	1.9100	0.7800
Viagens e Turismo	0.9800	0.8833	0.9300	23.5632	2.0900	1.0200
Exploração de Rodovias	0.9300	0.9400	0.9300	10.2802	1.9133	0.6150
Produtos de Uso Pessoal	0.9633	0.9067	0.9300	16.4444	2.4267	0.7400
Água e Saneamento	0.8900	0.9733	0.9300	4.2424	2.0667	1.0400
Energia Elétrica	0.9162	0.9557	0.9329	7.4514	2.1590	0.7700
Material Aeronáutico e de Defesa	0.9767	0.8967	0.9333	16.1765	2.0833	0.7500
Transporte Aéreo	0.9200	0.9500	0.9333	12.1951	2.9067	0.6700
Incorporações	0.8800	0.9967	0.9350	0.5464	1.9450	0.8200
Petroquímicos	0.9400	0.9400	0.9367	3.2787	2.3767	1.0000
Exploração de Imóveis	0.9056	0.9733	0.9367	4.2577	1.6022	0.7567

[1] Acessado em: <http://www.b3.com.br/pt_br/produtos-e-servicos/negociacao/renda-variavel/acoes/consultas/classificacao-setorial/>

Observa-se que as métricas de precisão não necessariamente refletem as métricas monetárias, pois o segmento de Eletrodomésticos conseguiu evitar a pior queda dentre os ramos de negócio mesmo mantendo F-Medida inferior a outros segmentos. O oposto pode ser identificado no segmento de Exploração de Imóveis, o qual mesmo atingindo boas métricas de desempenho, teve um resultado monetário modesto.

Independente do setor os algoritmos obtiveram bons resultados, com retornos financeiros variando entre 37.33% e 308%, nunca negativos, e perdas sofridas entre 0.54% e 43.67% das ocorridas no período. Mesmo nos setores onde houve menos perdas evitadas, como Papel e Celulose, Cervejas e Refrigerantes e Eletrodomésticos, a precisão das predições contribuiu para que as quedas fossem amenizadas por ganhos, conduzindo a resultados positivos apesar das perdas sofridas.

Em termos comparativos, o pior resultado obtido pelos algoritmos supera o melhor retorno alcançado pela estratégia *Buy & Hold*, respectivamente nos segmentos "Papel e Celulose" e "Minerais Metálicos".

5.2 Avaliação de padrões

Os resultados dos experimentos executados na seção anterior demonstram que, em concordância com estudos relacionados apresentados em 3, os algoritmos de aprendizagem de máquina atingem desempenho significativo ao antecipar movimentos nos preços baseados em notícias.

Esta seção, no entanto, busca analisar se os padrões obtidos apresentam coerência, proporcionando a racionalização dos resultados. A análise fornece indícios da confiabilidade das predições do modelo em cenários reais futuros.

Dos quatro melhores algoritmos avaliados na seção 5.1.6, três são considerados caixa preta por não proporcionarem interpretação direta de seus modelos, são eles: *MLP*, NB-G e *SVM*. O algoritmo RL, por outro lado, é considerado caixa branca por atribuir pesos aos atributos de acordo com sua importância na decisão da classe alvo, onde os pesos são, na verdade, o quociente entre a possibilidade do diagnóstico positivo em relação ao negativo, dado o atributo observado.

Os algoritmos caixa branca *KNN*, FA e AD, com grande potencial de interpretação de modelos, se mostraram incapazes de tratar a complexidade do problema na maioria dos casos testados. Para estes, foram escolhidas ações e datas específicas, realizando um estudo de caso a partir da explanação de instâncias.

Pelo esforço tanto computacional quanto na análise, a tentativa de interpretação dos modelos foi feita sobre um escopo reduzido. Em termos temporais, as análises se focam nos modelos usando o intervalo de publicação de D_{-1} e retornos em D_0, os quais atingiram

os melhores resultados. Para as ações foram escolhidos sobconjuntos menores, escolhidos dependendo da análise executada.

A Seção 5.2.1 inicia a avaliação de explanações de modelo, a partir da averiguação dos pesos atribuídos a cada atributo. Para a explanação de modelos foram escolhidos dois papéis, BRFS3.SA e PETR4.SA, que apresentaram grande repercussão na mídia nacional durante o período. A Seção 5.2.2 busca explanações de instância para predições em datas específicas, também para um conjunto limitado de ações.

5.2.1 Explanação de modelo

Explanação de modelo, ou explanação global, busca através de características do classificador explicar como as decisões, como um todo, são construídas. Para a mineração de textos na previsão quedas de preço, buscam-se as palavras mais importantes na decisão da direção do preço nos prazos futuros.

As análises a seguir se focam em avaliar os papéis que alcançaram maior repercussão no período analisado, a fim de verificar se palavras usadas na decisão do modelo estão relacionadas de alguma forma às quedas de preço ocorridas.

A primeira ação avaliada foi a BRFS3.SA, associada à empresa BRF, do segmento de carnes e derivados. Durante o período avaliado, houve várias publicações repercutindo investigações de envolvimento com corrupção e questionamentos sobre a qualidade dos produtos produzidos pela empresa.

As principais fases de investigação se deram entre março de 2017 e março de 2018. No mesmo período, houve notória queda de preço dos papéis da empresa, as quais foram antecipadas com certo sucesso por alguns dos modelos obtidos em 5.1.6.

O segundo papel avaliado é o PETR4.SA, associado à empresa de exploração de petróleo Petrobrás. Esta companhia também teve seu nome envolvido em diversas investigações de corrupção e casos de má administração entre 2015 e 2018, com nítida depreciação de suas ações.

As explanações de modelo são auferidas a partir da importância dos atributos associada às funcões de decisão, exibidas através de gráficos de nuvem de palavras.

5.2.1.1 Regressão Logística (RL)

No treinamento do modelo são estabelecidos pesos aos atributos de entrada baseando-se em quão significativos estes se mostram na separação entre as classes do atributo alvo. Os pesos são então aplicados de maneira linear à função de decisão, provendo um classificador capaz de classificar novos exemplos. A partir dos pesos estabelecidos na fase de treinamento, é possível investigar os atributos com maior peso na decisão da classe.

A Regressão Logística obteve resultados significativos para a maioria das ações avaliadas na seção 5.1.6. A Figura 27 exibe a nuvem de palavras composta pelos 400 termos de maior peso para decisão do modelo da ação BRFS3.SA, para predição da direção de retornos no prazo de 0 dias. As palavras mais significativas para o modelo alcançado para a ação PETR4.SA são exibidas no gráfico da Figura 28.

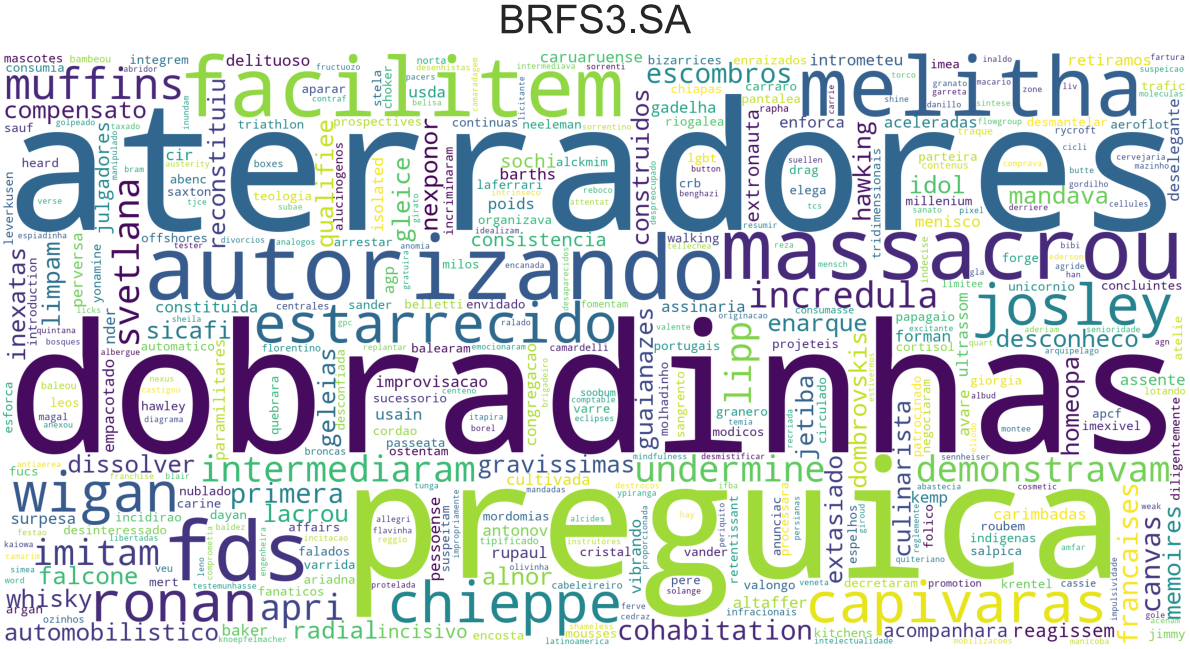


Figura 27 – Nuvem de palavras para classificador baseado em regressão logística, para o ativo BRFS3.SA.

Fonte: Produzido pelo autor com a biblioteca python WordCloud

Das palavras levantadas na Figura 27 é possível identificar o nome “josley”, semelhante ao nome “Joesley” cujo nome aparece envolvido em escândalos que abrangem a imagem da empresa BRFS3.SA. Além disso, notam-se várias palavras de conotação negativa, como “gravissimas” e “delituoso”. Para a PETR4.SA na Figura 28 são identificadas também palavras negativas como “sufocamento”, “engane”, “falidos” e “fiasco”.

Os modelos de regressão logística trabalham sempre sobre todo o conjunto de atributos, apenas discriminando pesos diferentes para cada campo. Assim, a decisão ocorre baseada não somente nos termos exibidos nos gráficos, mas também na ocorrência de cada uma das outras palavras do vocabulário.

Através do modelo de RL não é possível determinar a relação entre os atributos na tomada de decisão, apenas sua importância. Assim, os ofeitos de coocorrência de palavras não podem ser avaliados.

Figura 28 – Nuvem de palavras para classificador baseado em regressão logística, para o ativo PETR4.SA.

Fonte: Produzido pelo autor com a biblioteca python WordCloud

5.2.1.2 Modelos Simplificados

Dada a dificuldade em conduzir análises indiretas nos classificadores *MLP*, *SVM* e NB-G, a simplificação de modelos para torná-los mais explícitos se mostrou uma alternativa viável.

O algoritmo Naïve Bayes, em sua versão Gaussiana, não proporciona a possibilidade de vizualizar a importância dos atributos, pois a formulação matemática para características de entrada contínuas não envove o uso de pesos diretamente. A versão multinomial, por outro lado, tem seu uso indicado para problemas onde o domínio dos atributos é discreto, atribuindo pesos relativos a sua importância na separação das classes alvo.

De maneira semelhante, o algoritmo SVM com o uso de funções de *kernel* de base radial, usado nos testes da seção 5.1.6, possuem maior capacidade de modelar problemas cuja separação de classes exige recortes mais complexos, no entanto impossibilitam a verificação da importância dos atributos, pois estes são projetados em outro plano antes da elaboração da função de decisão. O uso de *kernels* lineares, apesar de limitados quanto a sua complexidade, permitem a observação direta da importância dos atributos na decisão dos classificadores alcançados.

A Tabela 24 exibe o comparativo de desempenho entre as versões otimizadas e simplificadas dos algoritmos, para publicações em D_{-1} e retornos em D_0. Para o algoritmo Naïve Bayes Multinomial o treino e teste dos modelos foi realizado na representação

simplificada *Bag of Words* (*BoW*), trocando a frenquência *TF-IDF* pelos valores binários 1 e 0, representando respectivamente a ocorrência ou não de uma palavra.

Tabela 24 – Comparativo entre versões otimizadas e simplificadas dos algoritmos SVM e Naïve Bayes.

Algoritmo	**Sigla**	***Precisão***	***Revocação***	***F-Medida***	***Retorno M.***	***Perda M.***
Naïve Bayes Gaussiano	NB-G	0.9268	0.8326	0.8643	1.5604	0.0840
Naïve Bayes Multinomial	NB-M	0.8686	0.6821	0.7264	1.4555	0.1416
SVM Kernel Radial	SVM-R	0.7163	0.6968	0.6696	1.2798	0.1685
SVM Kernel Linear	SVM-L	0.8461	0.7070	0.7332	1.3830	0.1511

Mesmo a Tabela 24 indicando degradação na medida revocação para o NB-M, ainda assim os resultados tanto para o NB-M quanto para o *SVM-L* apresentam bom desempenho nas predições.

Para a análise de termos as ações escolhidas foram novamente a BRFS3.SA e PETR4.SA, por sua repercussão no período. As figuras 29 e 30 exibem a núvens de palavras para os algoritmos NB-M e *SVM-L*, ambos avaliados para o papel BRFS3.SA.

Figura 29 – Nuvem de palavras para o algoritmo NB-M, avaliado para o papel BRFS3.SA.

Figura 30 – Nuvem de palavras para o algoritmo SVM-L, avaliado para o papel BRFS3.SA

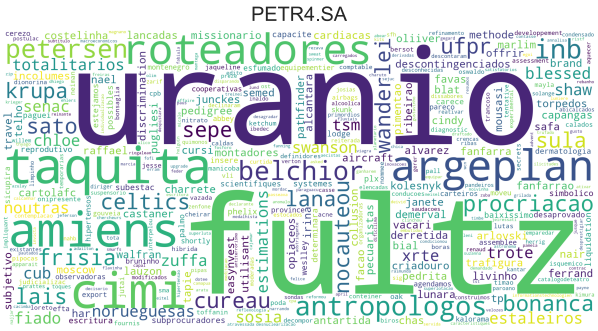
Fonte: Produzido pelo autor

As figuras 31 e 32 mostram as palavras mais importantes para os dois algoritmos, desta vez avaliados em relação ao papel PETR4.SA.

Para todos os casos novamente é difícil encontrar referências claras a termos relevantes para os papéis. Dentre as palavras levantadas, a mais evidente é apresentada na Figura 32, onde menciona-se o nome "argeplan", empreiteira associada a denúncias de corrupção.

Figura 31 – Nuvem de palavras para o algoritmo NB-M, avaliado para o papel PETR4.SA.

Figura 32 – Nuvem de palavras para o algoritmo SVM-L, avaliado para o papel PETR4.SA



Fonte: Produzido pelo autor

5.2.2 Explanação de instância

Diferente da explanação de modelo, a explanação de instância (ou explanação local) se restringe a obter do classificador elementos que expliquem apenas uma decisão, para apenas uma amostra. Não são analisadas todas as palavras mais importantes na decisão, mas aquelas que foram decisivas na classificação de um exemplo, uma data.

Para a investigação de explanações de instâncias foram avaliadas, no período de testes, as piores depreciações dentro do prazo de 0d, entre a abertura e o fechamento do mercado. As ações selecionadas e as predições efetuadas são exibidas na Tabela 25.

Tabela 25 – Piores retornos no período para ações selecionadas e predições.

Ação	Empresa	Data	Ret. d0	EMA	NB-G	KNN	AD	SVM	MLP	RL	FA
PETR3.SA	Petrobrás	2017-11-14	-0.0818	1	1	0	1	0	1	1	0
MGLU3.SA	Magazine Luiza	2017-10-17	-0.0888	1	1	0	1	1	1	1	0
BTOW3.SA	B2W	2017-11-07	-0.0952	1	1	0	1	1	1	1	1
USIM5.SA	Usiminas	2017-09-25	-0.1235	0	1	1	0	1	1	1	1
JBSS3.SA	JBS	2017-11-14	-0.0670	0	1	1	0	1	1	1	1
BRFS3.SA	Brasil Foods	2017-11-10	-0.0624	0	1	1	0	1	1	1	1

Dentre estes registros, para cada classificador foi escolhido um exemplar pelas seguintes regras:

1. O classificador previu corretamente a queda de preço.
2. Dentre as previsões corretas, seleciona-se a ação na qual o classificador obteve maior desempenho.

As explanações são exploradas em mais detalhes nas seções seguintes.

5.2.2.1 Árvore de Decisão (AD)

O conjunto de regras usado na decisão, uma vez estabelecido no treinamento do modelo, pode ser avaliado de maneira direta, inclusive graficamente no caso de classificadores de baixa complexidade.

Apesar de ajudarem a reduzir as perdas, nos experimentos da seção 5.1.6 este algoritmo obteve baixa taxa de sucesso ao antecipar quedas na maioria dos casos, eventualmente levando a resultados financeiros negativos.

Para esta análise foi escolhida a instância da ação BTOW3.SA, para o dia 7 de Novembro de 2017, no qual o classificador previu corretamente a queda. O caminho de decisão adotado pelo classificador para chegar ao rótulo é apresentado abaixo.

```
Regra 1: (habilidades = 0.94) >  0.14
Regra 2: (progression = 0.00) <= 0.68
Regra 3: (corpac      = 1.80) >  0.57
Regra 4: (filha       = 0.43) <= 0.45
Regra 5: (nisso       = 0.39) >  0.23
Regra 6: (slump       = 0.00) <= 1.06
```

Nota-se que a árvore apresenta um conjunto de regras simples para a decisão, mas claramente o encadeamento de regras não possui relação nem apresenta qualquer padrão relevante sobre a empresa. Para este caso pode-se concluir que a predição ocorre meramente levada pela entropia.

5.2.2.2 Floresta Aleatória (FA)

Para a avaliação do classificador obtido pela técnica Floresta Aleatória foi escolhido o papel JBSS3.SA, onde a técnica previou com sucesso a maior queda no período testado. Foi fornecida ao classificador a data 14 de novembro de 2017, e então foram investigados os caminhos de decisão que levaram à predição de queda. Como o algoritmo é na verdade um comitê de árvores de decisão, vários caminhos foram obtidos para definir a classe na predição, sendo o resultado definido por votação.

Dentre 2413 regras usadas na decisão, 9 se destacam por checarem a ocorrência dos nomes “geddel”, “aecio”, “neves” e “barroso”, envolvidos no processo de investigação de corrupção na empresa JBS. Além dos nomes as palavras “calote”, “sigilosos”, “multas”, “rescisões” e “investigados” também foram consideradas na decisão. As regras são apresentadas abaixo.

```
(geddel       = 1.38) > 0.86
(calote       = 1.69) > 1.15
(neves        = 0.67) > 0.40
(sigilosos    = 1.95) > 0.48
(multas       = 0.65) > 0.55
(rescisoes    = 1.86) > 0.93
(aecio        = 1.05) > 0.30
(investigados = 0.32) > 0.06
(barroso      = 0.66) > 0.58
```

Os termos "neves" e "investigados" aparecem ainda em outras regras, avaliados por outros limiares de decisão, mas apesar de ser possível identificar palavras pertinentes à empresa, o volume de regras e termos avaliados é alto, e na maioria dos casos os termos por si só não apresentam relação evidente com a empresa ou o negócio desta.

A análise de coocorrência de palavras através do caminho de decisão foi cogitada, mas entre os termos relevantes observados, todos apareciam encadeados em regras com palavras não diretamente relacionadas. Um exemplo é apresentado abaixo, com o termo "barroso".

```
Regra 1.  (itinerante = 1.07) <= 1.38
Regra 2.  (lutz       = 0.00) <= 0.88
Regra 3.  (licenciado = 1.22) <= 1.48
Regra 4.  (isentando  = 0.00) <= 1.16
Regra 5.  (barroso    = 0.66) >  0.58
Regra 6.  (delmiro    = 0.00) <= 1.00
Regra 7.  (desgasta   = 0.00) <= 0.97
Regra 8.  (colega     = 0.51) >  0.32
Regra 9.  (diriam     = 0.00) <= 0.90
Regra 10. (artistico  = 0.00) <= 1.44
Regra 11. (imbras     = 0.00) <= 1.54
Regra 12. (thayna     = 4.94) >  1.23
```

Os resultados apontam que os classificadores por FA fornecem indícios sobre o processo de decisão, contudo sua interpretação não é trivial.

5.2.2.3 K Vizinhos mais próximos (*KNN*)

O *KNN* obteve os piores resultados entre as técnicas de aprendizado de máquina no experimento do capítulo 5.1.6, mas teve analisado o caso específico do papel USIM5.SA, para a queda corretamente prevista para 25 de setembro de 2017.

Avaliando os classificadores obtidos e suas predições para outras ações, notou-se que os vizinhos averiguados eram frequentemente localizados no início do período de treino, no mês de agosto de 2016. Com o objetivo de amenizar o viés no cálculo de distâncias, foi realizado um experimento usando os valores de frequência *TF-IDF* discretizados, de maneira uniforme, em três faixas de frequências, entre os valores máximo e o mínimo

computados entre todos os termos. A ação avaliada foi a USIM5.SA, com retornos no prazo de 0 dias.

O classificador resultante obteve um retorno monetário de 113%, ante 72% da estratégia *Buy & Hold* e 59% para médias móveis exponenciais (EMA). Quanto a medidas de acerto, o modelo obteve 0.49 para precisão, 1.00 para revocação e 0.66 de F-Medida. Os vizinhos mais próximos obtidos para a data 25/09/2017 foram otimizados com $K = 15$, e são exibidos na Tabela 26.

Tabela 26 – Análise de vizinhos mais próximos para a data 25/09/2017, para o classificador USIM5.SA com retornos de 0 dias.

Vizinho	**Data**	**Retorno no dia**
1	2017-05-30	-0.004866
2	2017-06-20	-0.015152
3	2017-06-12	-0.015000
4	2017-05-15	0.040284
5	2017-07-07	-0.012220
6	2017-01-12	-0.002198
7	2017-06-05	0.010283
8	2017-05-08	-0.009662
9	2017-09-12	-0.001175
10	2017-07-03	0.012931
11	2017-06-26	0.035461
12	2016-10-07	0.013514
13	2017-05-17	-0.013453
14	2017-05-29	0.000000
15	2017-06-19	-0.005000

A Figura 33 exibe a evolução de preço para a ação no período, com a data de predição marcada com a linha pontilhada verical preta e os vizinhos definidos pelo *KNN* em cinza.

Do gráfico na Figura 33 nota-se que, exceto pelos dias 15 e 17 de julho de 2017, os vizinhos não foram escolhidos nas proximidades de quedas acentuadas, mas sim períodos estáveis e de alta. A Tabela 26 reforça essa observação, sendo que o maior retorno entre os vizinhos foi de 4% e o menor de -1.5%.

O *KNN* fornece uma interpretação muito útil ao investidor, permitindo que este tire suas próprias conclusões sobre as predições fornecidas. Pode-se analisar, por exemplo, se a situação atual de negociação realmente se assemelha às condições de mercado nas datas apontadas pelo algoritimo, através dos vizinhos.

Por outro lado, o algoritmo se mostra limitado se comparado a técnicas mais robustas como NB-G, SVM, RL e MLP, conduzindo a desempenhos modestos. Por este

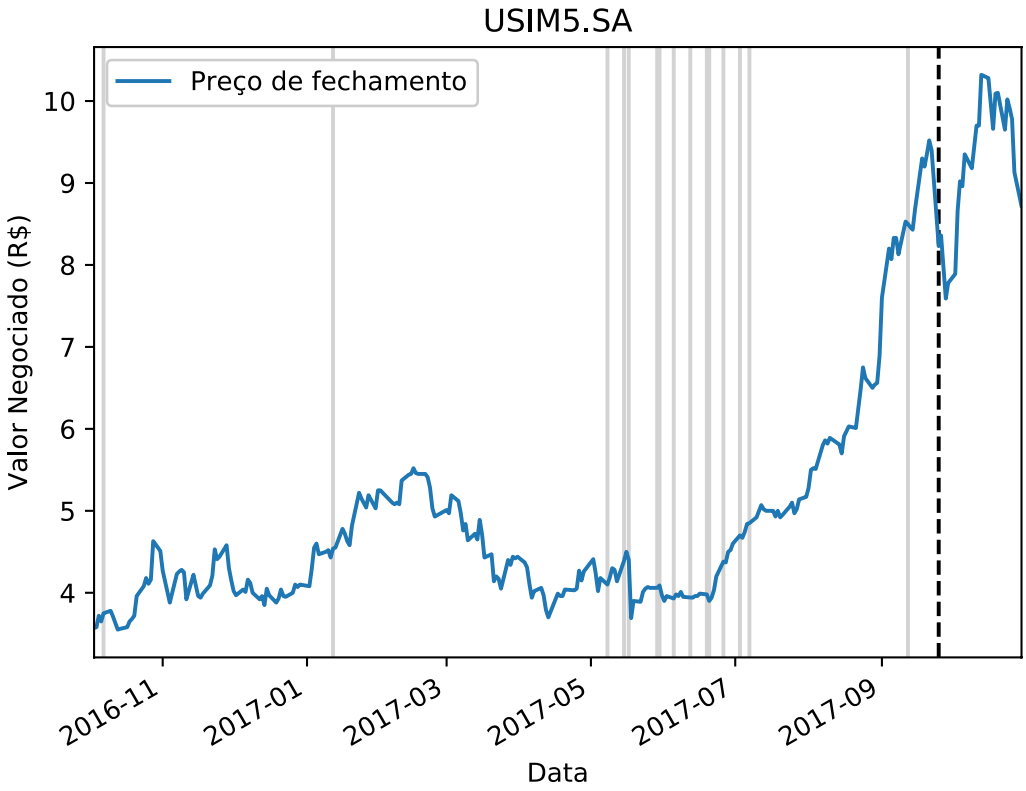


Figura 33 – Evolução de preços para o papel USIM5.SA, com a data prevista na linha vertical preta e vizinhos pelo *KNN* em cinza.

Fonte: Produzido pelo autor

motivo, seu uso se mostra restrito a poucos casos onde o modelo atinge resultados satisfatórios.

5.2.3 Discussão dos resultados

Pelas análises realizadas pode-se assumir as seguintes conclusões:

1. Árvores de Decisão, apesar de facilmente interpretáveis, não são adequadas ao tratamento do problema por modelarem funções de decisão demasiadamente simples. Florestas Aleatórias conseguem desempenho superior no tratamento do problema, no entanto elevam consideravelmente o esforço de análise para se obter explanações.

2. O algoritmo *KNN* possui um grande potencial em fornecer explanações simples e úteis, contudo seu uso é bastante limitado devido ao baixo desempenho na predição de depreciações.

3. As técnicas RL, NB-M e *SVM-L* atingem desempenho razoável e proporcionam a interpretação do modelo de maneira global. Contudo, a análise de palavras é dificultada pelo grande número de termos sem relação evidente com o negócio ou eventos associados a empresa.

4. De maneira geral, tanto explanações de instância quanto de modelos são viáveis, levantando a indícios dos temas usados pelos classificadores na tomada da decisão. Por outra lado, há um custo grande de esforço na análise manual dos termos pela complexidade dos modelos e alta dimensionalidade do vocabulário.

Conclusão

Esta dissertação investigou a relação entre as notícias públicas veiculadas na mídia brasileira e a queda de preços no mercado acionário nacional. Os objetivos foram atingidos, com os experimentos indicando forte relação entre as notícias e as cotações, sugerindo inclusive oportunidades de arbitragem em curto prazo.

Retomando as questões propostas em 5, baseado nos resultados dos experimentos é possível concluir:

Questão 1: Para o mercado de ações brasileiro, é possível evitar perdas com estratégias automáticas baseadas em mineração de texto de notícias em português?

Sim, é viável prever quedas de preço de ações brasileiras baseado em mineração de notícias em português. A seção 5.1.6 demonstra que as técnicas *MLP* e NB-G conseguiram evitar em média 90% das depreciações no período testado, fornecendo retornos acima de 200% do capital investido.

Questão 2: É possível obter dos classificadores explanações de modelo ou instâncias que descrevam os eventos que expliquem as depreciações?

Sim, é possível obter explanações dos classificadores, contudo a alta dimensionalidade do vocabulário e complexidade dos modelos com maior desempenho torna a análise de padrões não trivial. A explanação de instâncias através da análise de modelos como AD, FA e *KNN* fornece explanações mais diretas, no entanto com uso muito limitado devido ao baixo desempenho dos classificadores.

Comparados com as estratégias tradicionais *Buy and Hold* e Médias Móveis Exponenciais, os preditores fundamentados em aprendizado de máquina apresentaram resultados substancialmente melhores na maioria dos casos, com destaque para previsões de curto prazo com notícias do dia anterior, endossando os trabalhos semelhantes em mercados internacionais. Os algorimos foram especialmente exitosos ao evitar quedas, mas as técnicas Naive Bayes, *SVM*, *MLP* e Regressão Logística conduziram também a ganhos significativos de capital.

É demonstrado através de experimentos que há oportunidades significativas de ganho para predições de um dia, ainda consideráveis para o prazo de uma semana e modestas para um mês, sugerindo atraso na absorção de informações pelos investidores. Os testes indicam baixa eficiência informacional, pois os preços não se ajustam imediatamente assim que os dados são divulgados.

O uso de métricas financeiras foi aplicado tanto para treinamento dos modelos quanto para avaliação dos resultados. Apesar de não contribuírem como medidas de otimização na busca em *grid*, se mostraram relevantes na avaliação dos classificadores, pois métricas quantitativas não necessariamente implicacam em viabilidade financeira.

Em termos de verificação, nota-se que, apesar da importância da medida Revocação na identificação das quedas de preço, a Precisão dos classificadores conduz a melhores resultados monetários por impedir modelos com posições demasiadas conservadores. Em outras palavras, a redução de falso positivos leva não só a ganhos, mas sobretudo a recuperação de perdas sofridas em decorrência de falso negativos.

O uso de um corpo de notícias genéricas no treinamento do classificador se mostrou factível, sobretudo associada à seleção de atributos. Foi comprovado que o espaço de atributos pode ser reduzido a 2.5% de seu tamanho original com ganhos de desempenho, sendo F-Medida a melhor métrica para computar a importância.

Contribuições

Dentre as contribuições, apresentam-se:

1. Identificação de indícios da baixa eficiência informacional do mercado acionário brasileiro, através de experimentos que sugerem atraso no ajuste dos preços.

2. Criação de uma base de dados de notícias de cunho político e econômico, com granularidade diária, capturadas dos principais portais brasileiros[2].

3. Comparativo de estratégias de operação baseadas em aprendizagem de máquina com as tradicionais *Buy & Hold* e Médias Móveis Exponenciais, tendo como medidas de avaliação tanto métricas de desempenho quanto financeiras.

4. Proposta de abordagem preditiva de classificação de texto, com agrupamento de documentos em nível diário, corpo de notícias comum analisado individualmente sob o ponto de vista de cada ação, com vocabulário selecionado para cada ativo.

5. Análise de duas abordagens para a interpretação de padrões nas notícias:

    - A importância dos termos na determinação de classes, auferida pelos pesos nas funções de decisão dos classificadores;
    - A similaridade de contexto econômico entre datas, medida pela proximidade de frequência de termos.

[2] Disponível em <https://bit.ly/2mXUxix>.

Trabalhos futuros

Como trabalhos futuros, os seguintes tópicos podem ser explorados:

- **Otimização de portifólio:** Estender o estudo de classificadores, sobretudo de longo prazo, para seleção de ações mais promissoras para formação de portifólios.
- **Aprendizado online:** Explorar técnicas que permitam o treinamento contínuo dos modelos, garantindo que as predições sempre considerem os dados mais recentes disponíveis.
- **Aprendizado por transferência:** Dado o desempenho da rede *MLP*, pode-se estudar o treinamento de classificadores inicialmente por setor, e posteriormente especializar o modelo para ações específicas por *transfer learning*.
- **Custos de operação:** Relevar os custos transacionais na seleção de modelos, favorecendo aqueles que não só forneçam baixo risco e retornos atrativos, mas que o retorno financeiro líquido também seja viável.
- **Enriquecimento semântico:** Estudar métodos de enriquecimento semântico, tal como o tratamento de sinônimos. Estas abordagens devem promover tanto classificadores mais robustos quanto explanações de mais qualidade.

Referências

1 SILVA, R. M. *Da Navalha de Occam a um método de categorização de textos simples, eficiente e robusto*. 142 p. Tese (Doutorado) — Universidade Estadual de Campinas, 2017. Citado 5 vezes nas páginas 11, 32, 33, 35 e 84.

2 BILGINSOY, C. *A History of Financial Crises: Dreams and Follies of Expectations*. 1st. ed. New York: Routledge, 2014. 500 p. Citado na página 16.

3 WEINBERG, J. The Great Recession and its Aftermath. *Federal Reserve History*, nov 2013. Disponível em: <https://bit.ly/2UuCmvO>. Citado na página 16.

4 DE CAMARGOS, M. A.; BARBOSA, F. V. Teoria e evidência da eficiência informacional do mercado de capitais brasileiro. *Caderno de Pesquisas em Administração*, v. 10, n. 1, p. 41–55, 2003. ISSN 01036513. Citado 6 vezes nas páginas 16, 24, 26, 47, 48 e 52.

5 FAMA, E. F. Efficient Capital Markets - A Review of Theory and Empirical Work. *The Journal of Finance*, v. 25, n. 2, p. 383–417, 1970. ISSN 00221082. Citado 6 vezes nas páginas 16, 24, 25, 27, 46 e 48.

6 SHILLER, R. J. *Irrational Exuberance*. 3rd. ed. Princeton, New Jersey: Princeton University Press, 2000. 312 p. ISBN 978-0691050621. Citado 3 vezes nas páginas 16, 17 e 45.

7 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. ISSN 09574174. Citado 8 vezes nas páginas 16, 17, 24, 34, 45, 46, 50 e 52.

8 AASE, K.-G.; ÖZTÜRK, P. *Text Mining of News Articles for Stock Price Predictions*. 82 p. Tese (Mestrado) — Norwegian University of Science and Technology, 2011. Citado 3 vezes nas páginas 17, 45 e 50.

9 CHOWDHURY, S. G.; ROUTH, S.; CHAKRABARTI, S. News Analytics and Sentiment Analysis to Predict Stock Price Trends. *International Journal of Computer Science and Information Technologies*, v. 5, n. 3, p. 3595–3604, 2014. Citado 3 vezes nas páginas 17, 45 e 50.

10 KUMAR, B. S.; RAVI, V. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, v. 114, p. 128–147, 2016. ISSN 09507051. Citado 6 vezes nas páginas 17, 30, 45, 46, 50 e 52.

11 LOPES, T. J. P.; HIRATANI, G. K. L.; BARTH, F. J. Mineração de Opiniões aplicada à Análise de Investimentos. *WebMedia '08 Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, p. 117–120, 2008. Citado 4 vezes nas páginas 17, 48, 50 e 51.

12 REHBEIN, O. M. *Mineração de texto aplicado à análise de carteira de ações*. 96 p. Tese (Bacharelado) — Universidade de Santa Cruz do Sul, 2012. Citado 4 vezes nas páginas 17, 48, 50 e 51.

13 ALVES, D. S. *Uso de técnicas de Computação Social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores*. 117 p. Tese (Doutorado) — Universidade de Brasília, 2015. Citado 4 vezes nas páginas 17, 49, 50 e 51.

14 CORREIA, J. S. *Operando na bolsa de valores utilizando análise técnica*. 1. ed. São Paulo: Novatec, 2008. Citado na página 21.

15 PREETHI, G.; SANTHI, B. Stock market forecasting techniques: A survey. *Journal of Theoretical and Applied Information Technology*, 2012. ISSN 18173195. Citado na página 21.

16 LUENBERGER, D. G. *Investment Science*. [S.l.: s.n.], 1998. 1–20 p. ISSN 0018-9286. ISBN 9780195108095. Citado na página 22.

17 STRONG, N. Modelling Abnormal Returns: A Review Article. *Journal of Business Finance & Accounting*, v. 19, n. 4, p. 533–553, 1992. Citado 2 vezes nas páginas 23 e 24.

18 KAUFMAN, G. G.; SCOTT, K. E. What is systemic risk, and do bank regulators retard or contribute to it? *Independent Review*, v. 7, n. 3, p. 371–391, 2003. ISSN 10861653. Citado na página 24.

19 SANTOS, J. C. d. S. *Reação e Previsão: Tipificação do comportamento dos investidores no mercado de ações brasileiro*. 92 p. Tese (Livre Docência) — Universidade de São Paulo, 2017. Citado 3 vezes nas páginas 24, 25 e 48.

20 DE BONDT, W. F.; THALER, R. Does the Stock Market Overreact? *The Journal of Finance*, XL, n. 3, p. 793–805, 1985. Citado 2 vezes nas páginas 25 e 26.

21 BARBERIS, N.; THALER, R. A survey of behavioral finance. In: *Handbook of the Economics of Finance*. [S.l.]: Elsevier, 2003. v. 1, cap. 18, p. 1053–1128. ISBN 9780444513632. Citado na página 25.

22 SHILLER, R. J. Stock Prices and Social Dynamics. *Brookings Papers on Economic Activity*, v. 2, n. 1, p. 457–498, 1984. ISSN 0007-2303. Citado 2 vezes nas páginas 25 e 26.

23 LIN, T. C. W. A Behavioral Framework for Securities Risk A Behavioral Framework for Securities Risk. v. 325, 2011. Citado na página 26.

24 LI, J.; YU, J. Investor attention, psychological anchors, and stock return predictability. *Journal of Financial Economics*, v. 104, n. 2, p. 401–419, 2012. ISSN 0304405X. Citado na página 26.

25 HOLT, C. A.; LAURY, S. K. Risk aversion and incentive effects. *American Economic Review*, v. 92, n. 5, p. 1644–1655, 2002. ISSN 00028282. Citado 2 vezes nas páginas 26 e 51.

26 HALFELD, M.; TORRES, F. d. F. L. Finanças Comportamentais: aplicações no contexto brasileiro. *Revista de Administração de Empresas*, v. 41, n. 2, p. 64–71, 2001. Citado na página 26.

27 JEGADEESH, N.; TITMAN, S. Returns to Buying Winners and Selling Losers : Implications for Stock Market Efficiency. *The Journal of Finance*, XLVIII, n. 1, 1993. Citado na página 26.

28 GRAHAM, B.; DODD, D. *Security Analysis*. 6th. ed. [S.l.]: Mc Graw Hill, 2009. Citado na página 27.

29 LUENBERGER, D. G. *Investiment Science*. International edition. New York: Oxford University Press, 2009. Citado na página 27.

30 MINER, G. D.; ELDER, J.; NISBET, R. A. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. [S.l.: s.n.], 2012. ISBN 9780123869791. Citado na página 30.

31 LUHN, H. P. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, v. 2, n. 2, p. 159–165, 1958. ISSN 0018-8646. Citado na página 30.

32 WEISS, S. M. et al. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. [S.l.: s.n.], 2004. ISBN 9780203507223. Citado 2 vezes nas páginas 30 e 33.

33 BISHOP, C. M. *Pattern Recognition and Machine Learning*. 1. ed. New York: Springer, 2006. Citado 2 vezes nas páginas 30 e 67.

34 ABU-MOSTAFA, Y. S.; MAGDON-ISMAIL, M.; LIN, H.-T. *Learning From Data*. United States: AMLBook, 2012. Citado 6 vezes nas páginas 30, 31, 38, 39, 40 e 52.

35 MITCHELL, T. M. *Machine Learning*. 1. ed. New York: McGraw-Hill, 1997. Citado 5 vezes nas páginas 30, 36, 40, 41 e 42.

36 HOTHO, A.; NÜRNBERGER, A.; PAASS, G. A Brief Survey of Text Mining. *Ldv Forum*, v. 20, n. 1, p. 19–62, 2005. Citado 4 vezes nas páginas 31, 34, 35 e 41.

37 MAIMON, O.; ROKACH, L. *Data Mining and Knowledge Discovery Handbook*. 2. ed. Boston: Springer, 2005. ISBN 9780387699349. Citado 2 vezes nas páginas 31 e 35.

38 FORMAN, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, v. 3, p. 1289–1305, 2003. ISSN 15324435. Citado 2 vezes nas páginas 33 e 37.

39 ALMEIDA, T. A. et al. Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering. *Knowledge-Based Systems*, v. 108, p. 25–32, 2016. ISSN 09507051. Citado na página 34.

40 DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, v. 55, n. 10, p. 78, 2012. ISSN 00010782. Citado 2 vezes nas páginas 34 e 35.

41 FELDMAN, R.; SANGER, J. 1st. ed. New York: Cambridge University Press, 2007. 410 p. ISSN 14653133. ISBN 0521836573. Citado 2 vezes nas páginas 34 e 39.

42 TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, v. 37, p. 141–188, 2010. ISSN 10769757. Citado na página 34.

43 CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. *Computers and Electrical Engineering*, v. 40, n. 1, p. 16–28, 2014. ISSN 00457906. Citado 2 vezes nas páginas 35 e 38.

44 ZHENG, A.; CASARI, A. *Feature Engineering for Machine Learning and Data Analytics - Principles and Techniques for Data Scientists*. 1. ed. [S.l.]: O'Reilly, 2018. 263 p. ISBN 9781138744387. Citado na página 35.

45 YANG, Y.; PEDERSEN, J. O. A Comparative Study on Feature Selection in Text Categorization. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1997. p. 412—-420. Citado 2 vezes nas páginas 36 e 37.

46 MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. Citado 2 vezes nas páginas 36 e 37.

47 GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, v. 3, p. 1157–1182, 2003. Citado na página 38.

48 CHERKASSKY, V.; MULIER, F. *Learning From Data: Concepts, Theory and Methods*. 2. ed. New Jersey: Wiley, 2007. 557 p. Citado 3 vezes nas páginas 38, 39 e 40.

49 FACELI, K. et al. *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. 1. ed. Rio de Janeiro: LTC, 2011. 394 p. Citado na página 39.

50 COURONNÉ, R.; PROBST, P.; BOULESTEIX, A. L. Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, BMC Bioinformatics, v. 19, n. 1, p. 1–14, 2018. ISSN 14712105. Citado 2 vezes nas páginas 39 e 41.

51 LEWIS, R. J.; PH, D.; STREET, W. C. An Introduction to Classification and Regression Tree (CART) Analysis. *2000 Annual Meeting of the Society for Academic Emergency Medicine*, n. 310, p. 14p, 2000. Citado na página 41.

52 BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. [S.l.: s.n.], 2009. 479 p. ISSN 00992399. ISBN 9780596516499. Citado na página 41.

53 VELLIDO, A.; MARTIN-GUERRERO, J. D.; LISBOA, P. J. G. Making machine Learning models interpretable. *Proceedings ESANN*, v. 12, p. 163–172, 2012. ISSN 17476526 17476534. Citado na página 42.

54 MURDOCH, W. J. et al. Interpretable machine learning: definitions, methods, and applications. p. 1–11, 2019. Disponível em: <http://arxiv.org/abs/1901.04592>. Citado na página 42.

55 MOLNAR, C. *Interpretable Machine Learning*. [s.n.], 2019. ISBN 9781441907905. Disponível em: <https://christophm.github.io/Interpretable-Ml-Book/>. Citado 2 vezes nas páginas 42 e 43.

56 ROBNIK-SIKONJA, M.; BOHANEC, M. *Perturbation-Based Explanations of Prediction Models*. [S.l.]: Springer International Publishing, 2018. 159–175 p. Citado 2 vezes nas páginas 42 e 43.

57 AZAR, P. D. *Sentiment Analysis in Financial News*. 61 p. Tese (Bacharelado) — Harvard University, 2009. Citado 3 vezes nas páginas 45, 50 e 81.

58 ARAÚJO, J. G. de; MARINHO, L. B. Using Online Economic News to Predict Trends in Brazilian Stock Market Sectors. *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web - WebMedia '18*, p. 37–44, 2018. Citado 4 vezes nas páginas 45, 49, 50 e 51.

59 LI, X. et al. Does Summarization Help Stock Prediction? A News Impact Analysis. *IEEE Intelligent Systems*, v. 30, n. 3, p. 26–34, may 2015. ISSN 1541-1672. Citado 2 vezes nas páginas 45 e 50.

60 GUPTA, A. et al. Stock Market Prediction Using Data Mining. In: *International Conference on Advances in Science & Technology (ICAST)*. Maharashtra: [s.n.], 2019. v. 2, n. 2. ISBN 9781538619599. Citado 2 vezes nas páginas 45 e 50.

61 CARVALHO, V. P. de. *Previsão de séries temporais no mercado financeiro de ações com o uso de rede neural artificial*. 59 p. Tese (Mestrado) — Universidade Presbiteriana Mackenzie, 2018. Citado 4 vezes nas páginas 45, 48, 50 e 51.

62 AL-RUBAIEE, H. et al. Techniques for Improving the Labelling Process of Sentiment Analysis in the Saudi Stock Market. *International Journal of Advanced Computer Science and Applications*, v. 9, n. 3, 2018. ISSN 2158107X. Citado 2 vezes nas páginas 45 e 50.

63 SCHUMAKER, R. P. et al. Sentiment Analysis of Financial News Articles. *Communications of International Information Management Association*, p. 1–21, 2009. Citado 2 vezes nas páginas 45 e 50.

64 HADDI, E. *Sentiment Analisys: Text pre-procesing, reader views and cross domains*. Tese (Doutorado) — Brunel University London, 2015. Citado 3 vezes nas páginas 45, 50 e 81.

65 MAKREHCHI, M.; SHAH, S.; LIAO, W. Stock prediction using event-based sentiment analysis. In: *Proceedings - 2013 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2013*. [S.l.: s.n.], 2013. v. 1, p. 337–342. ISBN 9781479929023. Citado 2 vezes nas páginas 45 e 50.

66 MITTAL, a.; GOEL, A. Stock Prediction Using Twitter Sentiment Analysis. n. June, 2012. Citado 2 vezes nas páginas 45 e 50.

67 BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. *Journal of Computational Science*, Elsevier, v. 2, n. 1, p. 1–8, mar 2011. ISSN 1877-7503. Citado 2 vezes nas páginas 45 e 50.

68 YU, H. et al. Predictive ability and profitability of simple technical trading rules: Recent evidence from Southeast Asian stock markets. *International Review of Economics and Finance*, Elsevier Inc., v. 25, p. 356–371, 2013. ISSN 10590560. Citado 2 vezes nas páginas 45 e 50.

69 XING, F. Z.; CAMBRIA, E.; WELSCH, R. E. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, v. 50, n. 1, p. 49–73, 2018. ISSN 15737462. Citado 3 vezes nas páginas 45, 47 e 50.

70 LUCENA, P.; FIGUEIREDO, A. C. Pressupostos de eficiência de mercado: um estudo empírico na Bovespa. *GESTÃO.Org. Revista Eletrônica de Gestão Organizacional*, v. 2, n. 3, p. 1 – 13, 2004. ISSN 1679-1827. Citado 2 vezes nas páginas 48 e 52.

71 DE OLIVEIRA, F. A.; NOBRE, C. N.; ZÁRATE, L. E. Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index - Case study of PETR4, Petrobras, Brazil. *Expert Systems with Applications*, v. 40, n. 18, p. 7596–7606, 2013. ISSN 09574174. Citado 3 vezes nas páginas 48, 50 e 51.

72 YIM, J.; MITCHELL, H. A comparison of corporate distress prediction models in Brazil :. *Nova Economia Belo Horizonte*, v. 15, n. 1, p. 73–93, 2005. Citado 3 vezes nas páginas 48, 50 e 51.

73 KAHNEMAN, D.; TVERSKY, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, v. 47, n. March, p. 263–291, 1979. Citado na página 51.

APÊNDICE A – Caracterização de dados para publicações de D_{-1}

Na tabela abaixo é exibido o número de amostras de treino e teste para cada uma das 64 ações usadas no experimento da Seção 5.1.6. Para o ativo ABEV3.SA, por exemplo, foram usadas 419 amostras, das quais 281 foram aplicadas no treinamento do modelo e 138 no teste. A tabela mostra ainda o balanceamento dos dados, sendo que para a janela de 0d, a ação ABEV3.SA apresenta 155 amostras negativas ante 126 positivas para treino, e 75 negativas ante 63 positivas nos testes.

Para as janelas de publicações de D_{-5} e D_{-21} o balanceamento é semelhante, dado que as janelas de retornos futuros são os mesmos D_0, D_{+5} e D_{+21}. A variação acontece na quantidade de amostras, pois, pelo agrupamento de notícias, datas sem publicações são preenchidas com publicações de datas vizinhas, o que não acontece para D_{-1}. A quantidade total de amostras aumenta 5% e 10% para as janelas de D_{-5} e D_{-21}, respectivamente.

Ação	Total	Treino							Teste						
			Negativo			Positivo				Negativo			Positivo		
			0d	5d	21d	0d	5d	21d		0d	5d	21d	0d	5d	21d
ABEV3.SA	419	281	155	162	178	126	119	103	138	75	73	85	63	65	53
B3SA3.SA	419	281	159	175	187	122	106	94	138	76	70	65	62	68	73
BBAS3.SA	419	281	157	170	189	124	111	92	138	75	76	66	63	62	72
BBDC3.SA	419	281	162	169	175	119	112	106	138	71	63	60	67	75	78
BBDC4.SA	419	281	159	185	186	122	96	95	138	70	67	64	68	71	74
BBSE3.SA	419	281	135	140	152	146	141	129	138	68	63	52	70	75	86
BRAP4.SA	419	281	140	165	166	141	116	115	138	81	76	93	57	62	45
BRFS3.SA	419	281	133	129	124	148	152	157	138	57	37	21	81	101	117
BRKM5.SA	419	281	136	165	188	145	116	93	138	66	71	75	72	67	63
BRML3.SA	419	281	137	168	162	144	113	119	138	60	45	25	78	93	113
BTOW3.SA	419	281	147	163	164	134	118	117	138	72	75	85	66	63	53
CCRO3.SA	419	281	145	145	139	136	136	142	138	60	48	25	78	90	113
CIEL3.SA	419	281	132	121	91	149	160	190	138	68	64	70	70	74	68
CMIG4.SA	419	281	129	137	126	152	144	155	138	68	67	65	70	71	73
CPLE6.SA	419	281	134	141	146	147	140	135	138	64	57	59	74	81	79
CSAN3.SA	419	281	151	165	142	130	116	139	138	74	65	75	64	73	63
CSNA3.SA	419	281	130	137	155	151	144	126	138	62	70	50	76	68	88
CVCB3.SA	419	281	152	187	220	129	94	61	138	74	86	108	64	52	30
CYRE3.SA	419	281	143	167	169	138	114	112	138	70	69	67	68	69	71
ECOR3.SA	419	281	146	175	197	135	106	84	138	65	58	48	73	80	90
EGIE3.SA	419	281	126	139	134	155	142	147	138	63	61	76	75	77	62
ELET3.SA	419	281	127	113	132	154	168	149	138	62	63	52	76	75	86
ELET6.SA	419	281	131	127	134	150	154	147	138	61	62	50	77	76	88
EMBR3.SA	419	281	137	154	161	144	127	120	138	61	70	66	77	68	72
ENBR3.SA	419	281	146	152	171	135	129	110	138	64	57	38	74	81	100
EQTL3.SA	419	281	147	162	168	134	119	113	138	76	81	87	62	57	51
ESTC3.SA	419	281	159	166	169	122	115	112	138	78	66	82	60	72	56
FIBR3.SA	419	281	139	169	187	142	112	94	138	82	94	105	56	44	33

FLRY3.SA	419	281	146	158	183	135	123	98	138	65	70	51	73	68	87
GGBR4.SA	419	281	139	151	161	142	130	120	138	72	87	85	66	51	53
GOAU4.SA	419	281	146	150	169	135	131	112	138	69	84	85	69	54	53
GOLL4.SA	419	281	137	164	187	144	117	94	138	67	62	98	71	76	40
HYPE3.SA	419	281	142	156	177	139	125	104	138	77	74	80	61	64	58
IGTA3.SA	419	281	148	170	195	133	111	86	138	69	67	48	69	71	90
ITSA4.SA	419	281	153	173	177	128	108	104	138	64	76	76	74	62	62
ITUB4.SA	419	281	162	174	190	119	107	91	138	71	72	65	67	66	73
JBSS3.SA	419	281	136	149	142	145	132	139	138	66	70	77	72	68	61
KROT3.SA	419	281	145	158	180	136	123	101	138	63	43	18	75	95	120
LAME4.SA	419	281	128	150	158	153	131	123	138	72	64	70	66	74	68
LREN3.SA	419	281	159	173	207	122	108	74	138	71	73	48	67	65	90
MGLU3.SA	419	281	178	204	238	103	77	43	138	67	72	97	71	66	41
MRFG3.SA	419	281	148	153	164	133	128	117	138	62	49	58	76	89	80
MRVE3.SA	419	281	138	146	156	143	135	125	138	68	68	75	70	70	63
MULT3.SA	419	281	161	163	180	120	118	101	138	67	56	28	71	82	110
NATU3.SA	419	281	152	153	165	129	128	116	138	61	72	75	77	66	63
PCAR4.SA	419	281	138	169	188	143	112	93	138	67	62	59	71	76	79
PETR3.SA	419	281	133	152	148	148	129	133	138	82	94	96	56	44	42
PETR4.SA	419	281	140	153	157	141	128	124	138	74	94	91	64	44	47
QUAL3.SA	419	281	160	174	200	121	107	81	138	68	45	24	70	93	114
RADL3.SA	419	281	137	157	169	144	124	112	138	69	57	64	69	81	74
RAIL3.SA	419	281	156	179	177	125	102	104	138	75	76	74	63	62	64
RENT3.SA	419	281	154	180	185	127	101	96	138	70	82	108	68	56	30
SANB11.SA	419	281	153	166	182	128	115	99	138	75	81	104	63	57	34
SBSP3.SA	419	281	148	169	160	133	112	121	138	72	81	86	66	57	52
SMLS3.SA	419	281	155	170	193	126	111	88	138	70	70	54	68	68	84
SUZB3.SA	118	80	49	54	66	31	26	14	38	21	30	38	17	8	0
TIMP3.SA	419	281	158	175	191	123	106	90	138	75	79	96	63	59	42
UGPA3.SA	419	281	145	141	154	136	140	127	138	77	71	58	61	67	80
USIM5.SA	419	281	136	168	176	145	113	105	138	72	87	73	66	51	65
VALE3.SA	419	281	140	153	161	141	128	120	138	79	89	109	59	49	29
VIVT4.SA	419	281	153	162	144	128	119	137	138	70	64	54	68	74	84
VVAR3.SA	419	281	156	175	195	125	106	86	138	62	59	73	76	79	65
WEGE3.SA	419	281	152	149	192	129	132	89	138	60	79	70	78	59	68

Tabela 27 – Dados utilizados para o treinamento de cada um dos modelos. São apresentadas as quantidades de amostras de treino e teste, para cada uma das janelas de retorno.