I Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Predição do movimento do Índice Ibovespa a partir de notícias

Bruno Aparecido Barbosa[1], Ricardo Marcondes Marcacini[2], Solange Oliveira Rezende [3]
Instituto De Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)

1 Introdução

A predição do movimento do mercado acionário vem a muito tempo sendo o objetivo de pesquisadores e de agentes do mercado financeiro afim de antecipar movimentos para melhor alocação dos ativos na busca de evitar perdas e aumentar a rentabilidade nas aplicações financeiras.

Existe a defesa de que o movimento dos preços dos ativos se comporta de maneira aleatória, ou em função de muitas variáveis com previsão inviável [19]. Existem também estudos empíricos apontando contra exemplos onde há previsibilidades ao longo da história do mercado acionário. Tais estudos buscam padrões conhecidos chamados de anomalias de mercado, documentados e testados empiricamente, tais como os citados nos trabalhos [5, 21] direcionados ao mercado brasileiro.

A possibilidade de usar as informações passadas para predizer o movimento do mercado acionário traz a questão proposta por Eugene Fama em [8] quanto a análise da previsibilidade associada à eficiência informacional dos mercados. 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 [8], mercados mais eficientes se comportariam de maneira mais aleatória, e quanto maior a ineficiência, maiores as chances de sucesso em predições de arbitragem nos preços dos ativos, ou seja, existiria a possibilidade de encontrar padrões que se repetem no comportamento dos preços dos ativos.

A questão que se coloca então é se o mercado financeiro brasileiro apresenta eficiência informacional [8], ou seja, se os investidores dispondo de informações sobre as empresas interpretam racionalmente e de maneira instantânea essas informações. Supondo que sim, os preços dos ativos deveriam refletir as informações tão logo essas viessem a ser publicadas, evitando assim ganhos de arbitragem através dos preços dos ativos. No entanto, sabe-se que os preços atuais dos ativos

[1] brunobarbosa@usp.br
[2] ricardo.marcacini@icmc.usp.br
[3] solange@icmc.usp.br

podem não durar muito quando comparado ao atraso da publicação de eventos importantes e o tempo necessário para que uma nova notícia tenha sido assimilada pelos agentes financeiros [24].

Tem-se claro que a visão dos preços por parte do investidor se dá de uma forma puramente percebida, ao invés de algo relacionado diretamente aos custos de produção [19].

Sabe-se também que os jornalistas não apenas relatam o estado da realidade econômica dos ativos financeiros como também desempenham um papel ativo em sua criação de valor [29].

No atual cenário econômico brasileiro as condições vigentes tem incentivado cada vez mais investidores a optar por alocar seus recursos no mercado de renda variável (mercado acionário), visto por exemplo a taxa básica de juros da economia brasileira *SELIC* encontrasse muito próximo a sua mínima histórica (taxa essa tomada como livre de risco e balizadora de referência para qualquer investimento), tal cenário se reflete no aumento recorde de cadastro de novos investidores na bolsa de valores brasileira (*B3*) [3], tendo mais que dobrado em menos de um ano, atingindo quase 3 milhões de investidores.

Os novos entrantes no mercado acionário buscam informações sobre os ativos financeiros elevando a importância das informações para tomada de decisão como quanto e em quais ativos financeiros escolher e alocar seus recursos, o que também leva a cada vez mais meios de comunicação focarem na produção de conteúdo para esse novo público.

Ainda nesse contexto os meios digitais veem facilitando o acesso a informações com o crescimento constante de usuários da web em todo o mundo como também no Brasil [14]. Algo que também reforça a entrada de mais investidores no mercado acionário brasileiro é a facilidade que se tem de investir através de corretoras que buscam facilitar o acesso de novos investidores no mercado financeiro de ações.

O uso de algoritmos de aprendizagem de máquinas que consigam extrair conhecimento para insights ou mesmo precificar o valor futuro de ativos já é uma realidade na maioria dos mercados financeiros [12], seja através de *High Frequency Trading* (*HFT*), que são algoritmos de negociação que operam uma grande quantidade de ativos em um curtíssimo período de tempo, ou através de algoritmos que dão suporte a tomada de decisão através de previsões.

A aplicabilidade do uso de notícias financeiras (que é um dado não estruturado) em conjunto com algoritmos de aprendizado de máquinas para previsão do movimento ou preços do mercado financeiro já se mostrou eficaz em algumas pesquisas internacionais [6,11,13,16,18,19,22,23,25] como também em algumas pesquisas nacionais [1,7,10].

A proposta deste artigo é comparar a performance de duas técnicas de classificação de textos para a predição do movimento do mais importante índice de desempenho médio das cotações das ações negociadas na B3 (Brasil, Bolsa, Balcão) bolsa de valores oficial do Brasil, chamado Índice Ibovespa, formado pelas ações com maior volume negociados nos últimos meses. Serão considerados o método estado da arte para predição do movimento de ações através de notícias *BoW* (*Bag of Words*) com o classificador *Naive Bayes SVM* e o classificador BERT (*Bidirectional Encoder Representations from Transformers*) que é um modelo de linguagem baseado em redes neurais, com *text embedding* que vem atingindo bons desempenhos em classificação de textos [26].

Serão usadas notícias do caderno Mercado do jornal de maior circulação no Brasil, o jornal Folha de São Paulo segundo [9]. Além disso, serão selecionadas somente notícias que contenham o termo "Ibovespa" ou o nome das 10 principais empresas que compõem o Índice Ibovespa, e que juntas representam mais de *50%* da carteira do Índice de setembro de 2020 [4].

2 Fundamentos

As técnicas de mineração de texto vêm evoluindo gradualmente, as pesquisas com objetivo de prever o preço das ações usando dados textuais de notícias financeiras vem aumentando. A pesquisas apresentam constante progresso desde os anos 1990 apresentando uma dinâmica mais ativa desde os anos 2000, com progresso principalmente em aprendizagem de máquina [31]. Este capítulo resume os textos mais importantes nas pesquisas sobre mineração de textos e previsão do preço de ações com base em notícias financeiras.

2.1 Classificação de textos

A mineração de textos pode ser dividida em cinco etapas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento.

Na etapa inicial definimos o problema que pretendemos resolver, no nosso caso, temos como problema principal a questão se o modelo de linguagem baseado em redes neurais BERT apresenta melhor performance na tarefa de prever o movimento do mercado acionário brasileiro com notícias financeiras do que o método usualmente utilizado para esta tarefa com representação de atributos (*Bag-of-words*).

A segunda etapa consiste em obter uma representação estruturada apropriada para extrações de padrões das informações que pretendemos, será usado o método de representação *Bag-of-words* com classificador *Naive Bayes SVM* como *baseline* para com classificar BERT (*Bidirectional Encoder Representations from Transformers*) que é um método baseado em um modelo de *deep learning* chamado Transforme, desenvolvido pelo Google [26]. O BERT consiste em duas etapas de treinamento de uma rede neural: primeiro, o BERT é pré-treinado em grandes quantidades de texto, com um objetivo não supervisionado de modelagem de linguagem mascarada e previsão da próxima frase. Em seguida, essa rede pré-treinada e então ajustada em dados rotulados de tarefas específicas [26].

Após o uso do modelo de representação *Bag of words* há ainda a necessidade de seleção de um subconjunto de atributos relevantes para simplificação e evitar a maldição da dimensionalidade. Neste artigo, será usado o *Term Frequency-Inverse Document Frequency* (TF-IDF) para superar essa limitação. , como em [13].

A mineração de textos pode ser dividida em cinco etapas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento.

Na etapa inicial definimos o problema que pretendemos resolver com o método de mineração, bem como os critérios de avaliação que usaremos para compararmos com outras soluções ou mesmo avaliar a qualidade da solução. No nosso caso, temos como problema principal a questão se o modelo de linguagem baseado em redes neurais BERT apresenta melhor performance na tarefa de prever o movimento do mercado acionário brasileiro com notícias financeiras do que o método usualmente utilizado para esta tarefa com representação de atributos (*Bag-of-words*).

A segunda etapa consiste em obter uma representação estruturada apropriada para extrações de padrões das informações que pretendemos trabalhar [20], no nosso caso textos de notícias sobre o mercado financeiro. O abordagem mais popular usada é a *Bag-of-words* no campo de previsão de ações através de notícias financeiras [19]. Portanto será usado o método de representação *Bag-of-words* com classificador *Naive Bayes SVM* como *baseline* para com classificar BERT (*Bidirec-*

tional Encoder Representations from Transformers) que é um método baseado em um modelo de *deep learning* chamado Transforme, desenvolvido pelo Google [26]. O BERT consiste em duas etapas de treinamento de uma rede neural: primeiro, o BERT é pré-treinado em grandes quantidades de texto, com um objetivo não supervisionado de modelagem de linguagem mascarada e previsão da próxima frase. Em seguida, essa rede pré-treinada e então ajustada em dados rotulados de tarefas específicas [26].

Após o uso do modelo de representação *Bag of words* há ainda a necessidade de seleção de um subconjunto de atributos relevantes para simplificação e evitar a maldição da dimensionalidade. A seleção dos atributos é feita com base em dicionários predefinidos como usado em [27]. No entanto, o método *Bag of words* traz a desvantagem de ser difícil generalizar devido ao fato de que o conjunto de termos mudar com o tempo, e a seleção dos termos uma vez feita é estável. Este método pode ser baseado na ocorrência mínima por documento [22], ganho de informação [11] e qui-quadrado [13]. E por fim temos a representação dos atributos por um valor numérico para que possa ser inserido em algoritmo de aprendizado de máquina. O mais básico é uma representação binária que indica a ausência ou presença de um recurso. No entanto, esse método tem limitações porque não reflete a frequência da palavra, embora seja importante. Neste artigo, será usado o *Term Frequency-Inverse Document Frequency* (TF-IDF) para superar essa limitação, como em [13].

Na etapa de extração de padrões, que no consiste na predição do movimento do Índice Ibovespa foi utilizada co classificador *Naive Bayes SVM* que vem sendo utilizado em alguns artigos para essa tarefa com o método *Bag-of-words* [19, 22, 28] em comparação com o *fine-tuned* para classificação de documentos do BERT.

Por fim as etapas de pós-processamento e utilização do conhecimento são respectivamente as etapas onde avaliamos o conhecimento extraído na etapa de extração de padrões e em seguida buscamos entender seu funcionamento e usa-lo para auxílio a tomada de decisão segundo o problema proposto.

2.2 Trabalhos relacionados

George, Tony e Jin-Lung [30] analisaram a utilidade de notícias em conjunto com variáveis macroeconômicas para prever o retorno de ações setoriais no mercado de ações de Taiwan utilizando a abordagem de representação das notícias com o método TF-IDF com seleção de atributos em conjunto com o modelo SVM. Compararam os resultados com o mesmo modelo usando somente as variáveis macroeconômicas, sem as notícias. Chegaram a resultados em que a acurácia dos modelos que utilizam notícias tem uma performance superior do que os modelos que não utilizam notícias. Os autores também apontaram que para o período da crise de 2008 o efeito obtido com as notícias é menos efetivo, principalmente para notícias positivas e que o uso de uma única fonte de notícias também pode ter limitado os resultados da pesquisa.

Os autores Adm, Mahesan e Enrico [2] apontam que o uso de notícias é melhor utilizado para prever a volatilidade do retorno do ativo subjacente do que o movimento do ativo através de notícias, direcionando a possibilidade de maior aproveitamento do uso de notícias financeiras a maior potencialidade em contratos de derivativos de precificação, uma vez que esses tem seus preços alterados mais em virtude da volatilidade do ativo do que da direção do retorno. Seus experimentos utilizaram o índice de mercado NASDAQ dos Estados Unidos e ações de duas empresas, Goldman Sachs e J. P. Morgan em conjunto com arquivos de notícias da Reuters US. Utilizando o

modelo semântico para arquivos de texto e redução do números de atributos LDA (*Latent Dirichlet Allocation*) e predição com o modelo multinominal *Naïve Bayes*.

O autor Juvenal [7] abordou o uso de notícias para prever quedas no mercado acionário brasileiro através de notícias das ações que compõe o Índice Ibovespa, buscando além disso interpretabilidade dos resultados apontados pelos 7 experimentos testados. Os resultados apresentados apontam que as técnicas de mineração de textos se sobressaem às estratégias tradicionais de previsão de quedas de médias móveis e *Buy and Hold*, contudo a interoperabilidade dos modelos é limitada em função da complexidade dos classificadores e da alta dimensionalidade do vocabulário das notícias. Utilizou-se o método de extração de atributos TF-IDF com a seleção de atributos com os seguintes modelos Naïve Bayes, SVM, RNAs, Árvore de Decisão, Floresta Aleatória, Regressão Logística e K Vizinhos Mais Próximos com notícias de 10 veículos de comunicação diferentes.

Os autores Yauheniya, McGinnity, Sonya e Ammar [25] criaram a abordagem de usar várias categorias de notícias do setor de saúde com a técnica de MKL (Multiple Kernel Learning) e KNN vizinhos mais próximos alocando as notícias em cinco categorias com base em sua relevância para uma ação-alvo e seu sub-setor, indústria, indústria de grupo e setor. Cada categoria de notícias foi pré-processada independentemente das outras e cinco subconjuntos de dados foram formados. O MKL foi usado para aprender com esses novos subconjuntos, de forma que kernels independentes fossem utilizados para cada subconjunto. Diferentes tipos de kernel e várias combinações de kernel foram empregados. Os resultados obtidos mostraram que a maior precisão de previsão e retorno de negociação foram alcançados para MKL com cinco categorias de notícias utilizadas e dois kernels, polinomial e gaussiano, usados para cada categoria. Os autores pontuaram também uma possível direção de trabalho futuro é incluir fontes de dados adicionais, como preços históricos e fazer previsões com séries temporais e também que Kernels adicionais podem ser empregados para diferentes fontes de dados.

Os autores Xiaodong Li, Pangjing Wu e Wenpeng Wang [17] combinaram indicadores técnicos de preços de ações e análise de sentimentos de notícias criando um modelo de previsão que seja capaz de aprender informações sequenciais dentro de séries temporais de forma inteligente. Basicamente, foram convertido os preços históricos em indicadores técnicos que resumem os aspectos das informações de preços e modela sentimentos das notícias usando diferentes dicionários de sentimento e representam as notícias por vetores de sentimentos construindo uma rede neural LSTM (Long short-term memory) de duas camadas para aprender as informações sequenciais dentro de uma série de mercado. Tal abordagem supera o MKL e o SVM em precisão de previsão e pontuação F1-medida. Com melhor performance para o dicionário de sentimento de finanças (Loughran – McDonald Financial Dictionary). Os autores pontuaram precisariam mudar a análise de sentimento de texto completo apara análise de sentimento baseada em eventos.

A abordagem que será utilizada neste artigo utiliza o método *Bag of Words* com o classificador *Naive Bayes SVM* em comparação com o classificador BERT. Utilizando para ambos os classificadores o mesmo grupo de notícias (dataset) treinando os classificadores com 70 % das notícias ordenadas em ordem cronológica e depois testando o resultado dos modelos com os 30 % de notícias restantes comparando-se métricas de avaliação dos modelos como também expondo o entendimento que pode-se tirar dos resultados obtidos.

3 Abordagem proposta

A Figura 1 apresenta as etapas que compõem a construção dos modelos preditivos.

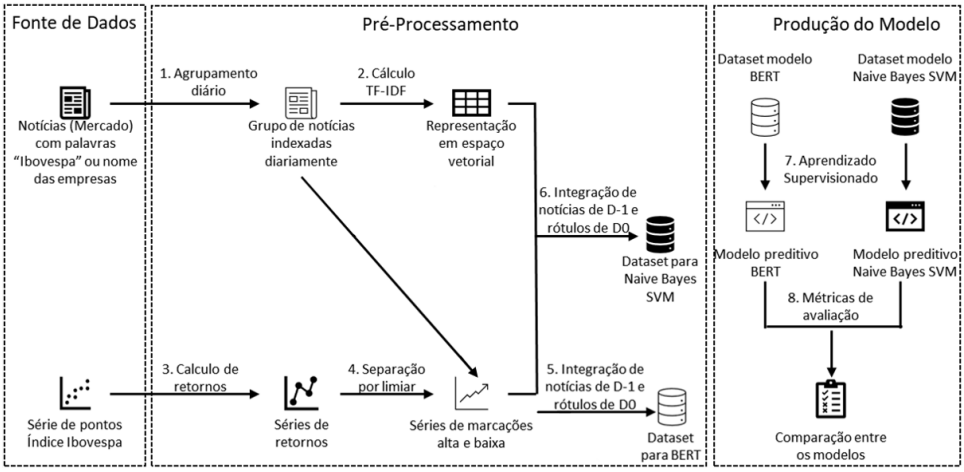

Figura 1: Etapas para construção do modelo preditivo

A primeira etapa consiste em agrupar as notícias e indexá-las por data. Serão agrupadas todos as notícias que foram publicadas em um mesmo dia útil e para fins as notícias públicas em finais de semanas ou feriados serão consideridados o último dia útil antes da data de publicação.

O segundo passo trata todas as 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.

Considerando o Índice Ibovespa e seu retorno diário, as etapas 3 e 4 calculam o retorno diário e os transformam em rótulos de alta ou baixa. 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 5 as linhas do agrupamento de notícias indexadas diariamente são integradas à série de rótulos, pela data, tal que cada linha do dataset integrado apresenta as notícias indexadas para D_0 e os retornos no prazo futuro, para D_1. Esse será o dataset usado para treinar e testar o modelo BERT.

No passo 6 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_0 e os retornos no prazo futuro, para D_1 O dataset produzido possui alta dimensionalidade devido ao grande volume de notícias. Esse será o dataset usado para treinar e testar com o classificador Naive Bayes SVM.

No passo 7, as amostras são divididas entre os conjuntos de treino e teste e aplicadas na produção para ambos os modelos. O conjunto de treino é aplicado ao aprendizado supervisionado, o conjunto de testes é usado para produzir métricas de avaliação da qualidade do modelo.

Por fim no passo 8, as métricas apuradas entre os modelos são capturadas e comparadas entre acuraria de F1-Medida.

4 Avaliação experimental

A seguir são apresentados o conjunto de dados utilizados, a definição dos intervalos temporais e medidas de avaliação usadas para avaliar a proposta.

4.1 Conjunto de Dados

Foram consideradas notícias do jornal Folha de São Paulo do Caderno Mercado. Caderno esse que busca trazer as principais decisões políticas, econômicas, de grandes negócios nacionais e globais, como também as principais alterações com reflexos no Brasil.

O banco de dados utilizado foi disponibilizado pelo site Kaggle [15], contendo 167.053 notícias crawliadas diretamente do site do jornal no período de janeiro de 2015 até setembro de 2017 contendo títulos, url das notícias, texto completo das notícias publicadas, data de publicação de cada notícia e o caderno de origem de cada notícias. Para esse experimento utilizou-se somente notícias a partir de 01 de janeiro de 2016, visto o custo de processamento dos modelos.

Na figura 2 temos o número de notícias consideradas nesse experimento por mês.

Para fins de teste foram consideradas 70% dos dias do intervalo de notícias consideradas e 30% dos dias para validação dos modelos. Sendo que como o dataset de treino com a proporção de 70% considera notícias até a data 17 de fevereiro de 2017 e o de validação inicia-se após essa data. Com essa separação temos 412 dias de notícias para treino e 176 para validação, lembrando que cada dia contém todas as notícias publicadas naquele dia.

Para esse experimento considerou-se somente notícias contendo palavras com o nome das 10 maiores empresas que compõem o índice Ibovespa como também o termo IBOVESPA, visto que o caderno Mercado traz informações não relacionadas somente ao mercado financeiro, como também matérias sobre planejamento financeiro, situação econômica de cidades entre outros temas relacionados a finanças.

Para esse experimento considerou-se somente notícias contendo palavras apresentadas na tabela 4.1, visto que o caderno Mercado traz informações não relacionadas somente ao mercado financeiro, como também matérias sobre planejamento financeiro, situação econômica de cidades entre outros temas relacionados a finanças.

Referência	Nome Busca
Índice Ibovespa	Ibovespa
Petróleo Brasileiro	Petrobras
Banco Itaú	Itaú
Banco Bradesco	Bradesco
Companhia Vale do Rio Doce	Vale
AMBEV	Ambev
Magazine Luiza	Magazine Luiza
Banco do Brasil	Banco do Brasil
Itaúsa Investimentos Itaú	Itaúsa
BRASIL, BOLSA, BALCÃO	B3
WEG	WEG

4.1

Além disso, foram consideras somente notícias a partir de *jan/2016* visto o custo computacional de processamento dos modelos e seleção de features. O volume mensal das notícias usadas nos experimentos são apresentadas na figura 2

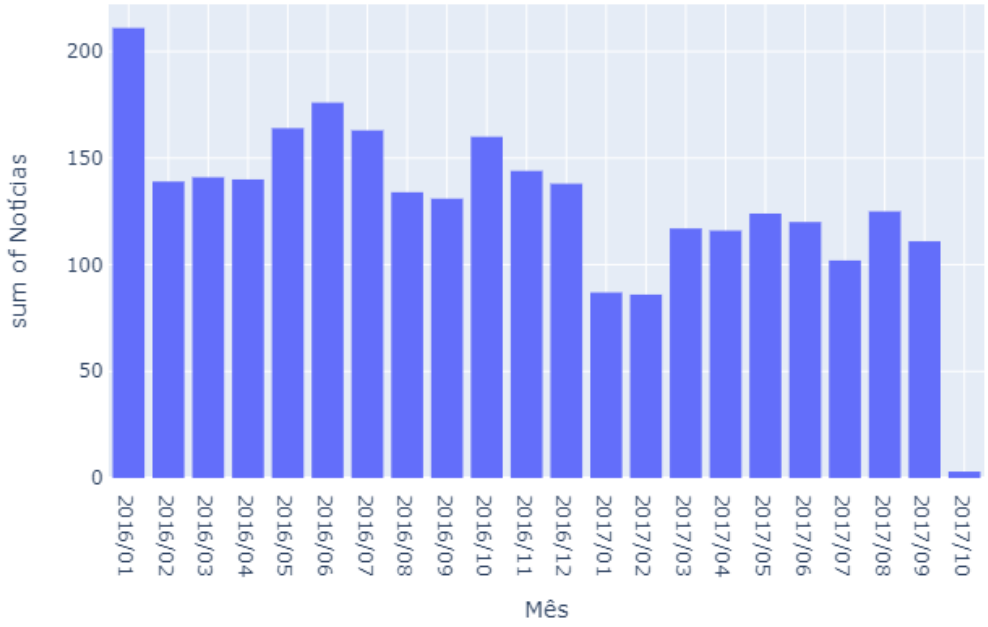

Figura 2: Número de noticias consideradas por mês

Foi utilizado o provedor Yahoo para coleta das séries quantitativas dos pontos do Índice Ibovespa por facilitar a captura automática e por fornecer os valores de abertura e fechamento do mercado.

Foi considerado quedas qualquer valores retorno do índice abaixo de 0 e altas qualquer valores de retorno acima de 0. Para o período analisado a serie apresenta dados razoavelmente balanceados retornos de alta e baixa do índice sendo 325 (60%) dias de alta e 263 (40%) dias de baixa. Abaixo, na figura 3, podemos ver a evolução do Índice Ibovespa no período, onde as linhas verdes verticais simbolizam dias de fechamento em alta e retas verticais em vermelho dias de baixa. Sendo que para dataset de treino temos 205 (54%) dias de alta e 178 (36%) de baixa.

Nota-se uma tendência de elevação, o que explica a menor quantidade de exemplos de alta.

4.2 Intervalos Temporais

Para fins da previsão de alta ou baixa no Índice Ibovespa o tempo das observações, tanto das notícias quanto dos pontos do índice são fundamentais na definição do objetivo do classificador.

Foram consideradas que dado uma notícia em um dia D_0, tenta-se inferir se a variação dos pontos do Índice entre a abertura e o fechamento do mercado em um dia útil posterior D_1. Com

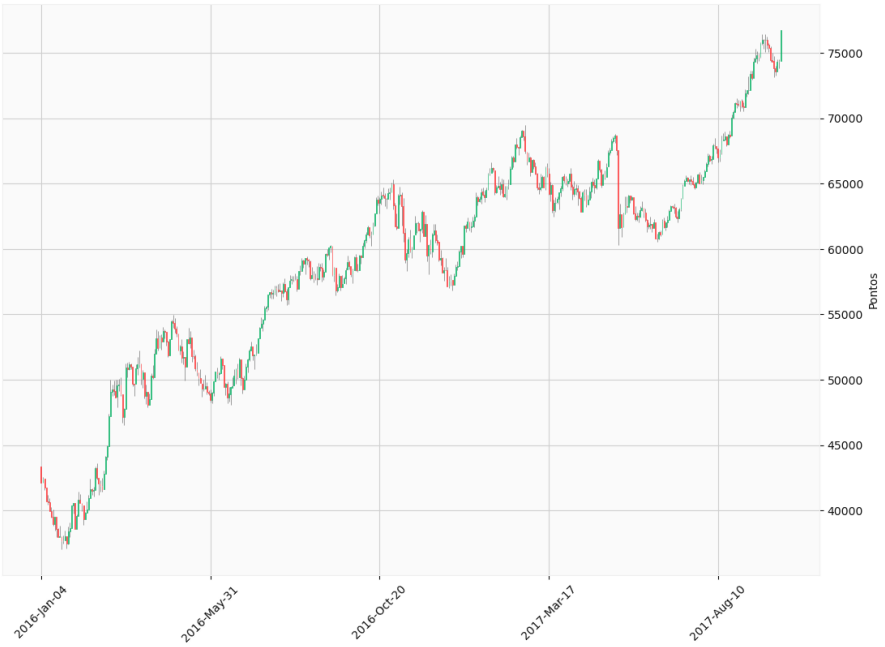

Figura 3: Histórico de pontos do Índice Ibovespa ajustado

isso temos duas relações de ajuste ao tempo, a das notícias usadas na predição e a segunda dos retornos.

O experimento foi executado tendo em vista dias úteis para as notícias, e para as notícias publicadas em finais de semana e feriados foram associadas ao dia útil imediatamente anterior, e para os cálculos dos retornos também foram consideras somente dias úteis.

4.3 Medidas de Avaliação

As medidas de avaliação são usadas para mensurar a qualidade dos modelos aplicadas sempre sobre aos dados de teste ou validação.

Nesse trabalho serão consideradas as medidas *Revocação* que no nosso contexto representa a quantidade de previsões de alta e baixa previstas dentre as ocorridas, *Precisão* quantidade de previsões de alta e queda realmente ocorreram e *F-Medida* como peso igual tanto para *Revocação* quanto para *Precisão* na forma de *F-medida* harmônica, promovendo o equilíbrio entre altas e quedas.

5 Resultados

Este capítulo discute os resultados do sistema de previsão de movimento de ações baseado em notícias proposto.

E realizada uma análise comparativa dos dois métodos propostos. Comparamos a precisão, revocação e a F1 da previsão do movimento das ações.

A Tabela 1 mostra os resultados experimentais obtidos a partir da comparação entre os dois modelos propostos. A primeira coluna da Tabela 1 mostra os resultados produzidos quando o método proposto é utilizado (BERT). A segunda coluna representa os resultados com o método mais comumente utilizado para predição do movimento de ações com base em notícias.

Tabela 1: Resultados

Métrica	BERT	*Naive Bayes SVM*
Revocação	0.48	0.51
Precisão	0.47	0.51
F-medida	0.54	0.51

Os resultados mostram que o método proposto BERT apresenta melhores resultados na métrica de avaliação *F-Medida* que o método do estado da arte. No entanto, para as duas demais métricas o modelo proposta apresenta resultados piores. Os resultados sugerem que considerar a modelos de linguagem baseados em redes neurais para predição do movimento do mercado acionário podem apresentar resultados equiparável aos modelos comumente utilizados.

6 Conclusão e trabalhos futuros

A previsão do preço das ações usando aprendizado de máquina tem sido estudada ativamente, e pesquisas sobre a previsão dos preços das ações com base em dados não estruturados têm atraído considerável atenção. Como é impossível para os investidores lerem todas as notícias sobre ações, os investidores podem obter benefícios potenciais usando sistemas automatizados que podem identificar informações de várias fontes e prever com precisão as mudanças nos preços de mercado.

Neste estudo, consideramos usar o modelo BERT em comparação com o estado para arte para a previsão do movimento do Índice Ibovespa utilizando notícias do mercado financeiro de uma única fonte, essa provavelmente é uma limitação desse estudo, uma vez que fontes diversas podem trazer analises diferentes e com vieses e perspectivas diferentes quanto as empresas mais importantes do índice.

Outro ponto de possível melhoria para estudos futuros é considerar *embbedings* pré-treinadas para o contexto de mercado financeiro, uma vez que tal estratégia elevaria o nível semântico para compreensão de informações do mercado financeiro pelas redes neurais.

Além disso, valeria a comparação com outros métodos de representação semânticas estado da arte para esta mesma tarefa, ou mesmo pré treinar uma *embbeding* com as notícias que estão sendo usadas para classificação provavelmente elevariam a performance do classificador.

Agradecimentos

Os autores agradecem os auxílios fornecidos para o desenvolvimento deste trabalho, FAPESP (2019/25010-5 e 2019/07665-4) e CNPq (426663/2018-7).

Referências

[1] Alves, D. S. (2015). Uso de técnicas de computação social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores. *Tese de Doutorado*.

[2] Atkins, A., Niranjan, M., and Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, 4(2):120–137.

[3] B3 (2020a). Histórico pessoas físicas (`https://comoinvestir.thecap.com.br/numero-de-investidores-pessoa-fisica-na-b3-b3sa3-bate-recorde/`). Acessado em 22/11/2020.

[4] B3(2020a). Índice Bovespa (Ibovespa) (`http://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/indice-ibovespa-ibovespa-composicao-da-carteira/` Acessado em 22/11/2020.

[5] Camargos, M. A. d. and Barbosa, F. V. (2010). Teoria e evidência da eficiência informacional do mercado de capitais brasileiro.. *REGE Revista de Gestão*, 10(1).

[6] Chowdhury, S. G., Routh, S., and Chakrabarti, S. (2014). News analytics and sentiment analysis to predict stock price trends. *International Journal of Computer Science and Information Technologies*, 5(3):3595–3604.

[7] Duarte, J. J. et al. (2019). Influência dos textos de notícias na queda de preços no mercado de ações brasileiro. *Tese de Mestrado*.

[8] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance* , 25(2):383–417.

[9] Folha de São Paulo (2021a). Com crescimento digital, Folha lidera circulação total entre jornais brasileiros

(`https://www1.folha.uol.com.br/poder/2020/01/folha-cresce-e-lidera-circulacao-entre-jornais-do-pais-em-2019.shtml`

[10] FROZZA, R. and REHBEIN, O. (2012). Mineração de texto aplicado a análise de carteira de ações. *Anais do Salão de Ensino e de Extensão*, page 32.

[11] Groth, S. S. and Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, 50(4):680–691.

[12] Gurav, U. and Sidnal, N. (2018). Predict stock market behavior: Role of machine learning algorithms. In *Intelligent Computing and Information and Communication*, pages 383–394. Springer.

[13] Hagenau, M., Liebmann, M., and Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *DecisionSupport Systems*, 55(3):685–697.

[14] Internet World Stats (2121a). South America (`https://www.internetworldstats.com/south.htm#br/`). Acessado em 22/11/2020.

[15] Kaggle (2020). News of the brazilian newspaper. (`https://www.kaggle.com/marlesson/news-of-the-site-folhauol/`) Acessado em 22/11/2020.

[16] Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., and Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, 278:826–840.

[17] Li, X., Wu, P., and Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of hong kong. *Information Processing & Management*,page 102212.

[18] Nam, K. and Seong, N. (2019). Financial news-based stock movement prediction usingcausality analysis of influence in the korean stock market. *Decision Support Systems*,117:100–112.

[19] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

[20] Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.

[21] Santos, J. C. d. S. *Reação e previsão: tipificação do comportamento dos investidores no mercado de ações brasileiro*. PhD thesis, Universidade de São Paulo.

[22] Schumaker, R. P. and Chen, H. (2009). A quantitative stock prediction system based on financial news. *Information Processing & Management*, 45(5):571–583.

[23] Shen, S., Jiang, H., and Zhang, T. (2012). Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford,CA*, pages 1–5.

[24] Shiller, R. J. (2015) *Irrational exuberance: Revised and expanded third edition*. Princeton university press.

[25] Shynkevich, Y., McGinnity, T. M., Coleman, S. A., and Belatreche, A. (2016). Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decision Support Systems* , 85:74–83.

[26] Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages194–206. Springer.

[27] Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information? *The Review of Financial Studies*, 24(5):1481–1512.

[28] Wang, S. I. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94.

[29] Wisniewski, T. P. and Lambe, B. (2013). The role of media in the credit crunch: The case of the banking sector. *Journal of Economic Behavior & Organization*, 85:163–175.

[30] Wu, G. G.-R., Hou, T. C.-T., and Lin, J.-L. (2019). Can economic news predict taiwan stock market returns? *Asia Pacific management review*, 24(1):54–59.

[31] Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., and Zhang, J. (1998).Daily stock market forecast from textual web data. InSMC'98 *Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat.No. 98CH36218)*, volume 3, pages 2720–2725. IEEE.