Sentiment Analysis Applied to News from the Brazilian Stock Market

Brenda A. Januário, Arthur E. de O. Carosia, Ana E. A. da Silva, Guilherme P. Coelho, *Senior Member, IEEE*

***Abstract*—Investments in the stock market have grown in Brazil in recent years, especially considering the individual number of investors. According to data from April 2020, the Brazilian stock market reached the historic mark of 2.38 million active investors, and with this scenario, there is an increasing need to study the Brazilian financial market, seeking to better understand its fluctuations. Recent work in the literature indicates that a company's stock values can be influenced by published news. Therefore, this work contributes to the automatic sentiment analysis applied to news written in Portuguese and related to the Brazilian stock market. For this, we performed three sentiment analysis strategies: two based on machine learning, using the Naive Bayes classifier and a Multilayer Perceptron neural network; and the other based on the lexical approach. Also, we proposed two dictionaries, focused on the financial domain and adapted to Portuguese. Our results show that the Naive Bayes classifier and the Multilayer Perceptron overcomes the best lexical approach. It is worth mentioning that the accuracy achieved by the best lexical approach was with the adapted dictionary proposed here.**

***Index Terms*—Sentiment Analysis, Text Mining, Stock Market.**

I. Introdução

O mercado de ações, ambiente no qual empresas de capital aberto negociam frações de seu patrimônio, tem feito cada vez mais parte das opções de investimento dos brasileiros. A bolsa brasileira, B3 (antiga Bovespa), atingiu em abril de 2020 a marca histórica de 2,38 milhões de investidores ativos [1]. Com o crescimento dessa demanda, há uma necessidade cada vez maior de se estudar o mercado financeiro brasileiro, buscando melhor compreender suas oscilações.

Uma das maneiras de analisar as mudanças no valor das ações e verificar possíveis associações é por meio de dados publicados *online*. A análise da opinião pública está gerando crescente interesse da comunidade científica e do mundo dos negócios, uma vez que é possível extrair informações desses dados e utilizá-los em estratégias para empresas, movimentos políticos, campanhas de marketing e predição da bolsa de valores [2].

Trabalhos recentes na literatura científica já averiguaram que, no mercado moderno de ações, notícias podem influenciar o preço das ações na bolsa, ainda que de forma rápida e puramente especulativa [3]. O valor de cada ação é determinado no momento da compra e venda pelo investidor que negocia, logo a opinião global expressa nas notícias consumidas no momento do investimento pode influenciar no valor de um ativo no mercado. Assim, ideias e opiniões manifestadas nestas notícias podem ser identificadas automaticamente através de processos computacionais [4].

Para isso, a Análise de Sentimentos (AS) é uma técnica de Processamento de Linguagem Natural (PLN) que busca analisar opiniões, sentimentos e emoções presentes em dados não-estruturados [5]. Existem duas técnicas principais para o problema de extração de sentimentos em textos: (*i*) abordagens baseadas em Aprendizado de Máquina (AM) e (*ii*) abordagens lexicais [6]. A primeira delas baseia-se na capacidade do sistema em aprender automaticamente com a experiência. Já a segunda abordagem é baseada em tratamentos léxicos por meio de dicionários, que envolvem o cálculo da polaridade de um texto a partir da orientação semântica das palavras contidas nele [7].

Dessa forma, o presente trabalho tem como objetivo extrair informações de notícias pertencentes ao domínio financeiro brasileiro, publicadas online, utilizando AS. Para isso, propõe-se um estudo comparativo entre as duas estratégias de análise de sentimentos: (1) aprendizado de máquina, por meio dos algoritmos *Naive Bayes* e *Multi-layer Perceptron* (MLP), que estão entre os mais utilizados para esta tarefa [8]; e (2) abordagem lexical, por meio de dois dicionários léxicos, técnica que tem apresentado bom desempenho na literatura para prever sentimentos em domínios específicos [9], [10].

Este trabalho também apresenta a criação de dois dicionários voltados para área de finanças em Português, além de apresentar uma estratégia de combinações de diferentes dicionários para AS de textos dessa área, contribuições que podem ser utilizadas em trabalhos futuros. Por fim, este trabalho contribui para o estudo de estratégias de PLN voltadas ao mercado brasileiro, que apresenta poucos trabalhos publicados até o momento [4].

O restante deste trabalho está organizado da seguinte forma: a Seção 2 aborda os conceitos teóricos necessários para o desenvolvimento do trabalho; a Seção 3 detalha as metodologias que foram utilizadas em cada etapa do estudo; a Seção 4 descreve e analisa os resultados obtidos; e, por fim, a Seção 5 conclui o trabalho.

II. Fundamentação Teórica

A técnica de Análise de Sentimentos (AS) visa identificar como os sentimentos são expressos em textos e se as expressões indicam opiniões positivas (favoráveis) ou negativas (desfavoráveis) em relação a um determinado assunto. Assim,

Carosia, A. E. O. is with the Federal Institute of São Paulo (IFSP), São João da Boa Vista - SP - Brazil and with the School of Technology (FT), University of Campinas (UNICAMP), Limeira - SP - Brazil. e-mail:arthuremanuel.carosia@ifsp.edu.br

Silva, A. E. A.; Januário, B. A.; Coelho, G. P. are with the School of Technology (FT), University of Campinas (UNICAMP), Limeira - SP - Brazil. e-mails: aeasilva@unicamp.br; brenda_janu@icloud.com; gpcoelho@unicamp.br

a AS envolve a identificação de expressões de sentimento, polaridade e força das expressões, e sua relação com o assunto [11]. O principal objetivo da AS é definir técnicas automáticas capazes de extrair informações de textos em linguagem natural, a fim de criar conhecimento estruturado que possa ser utilizado por um sistema de apoio ou tomador de decisão [12].

A seguir, são apresentadas as formas mais comuns de realizar AS na literatura: abordagem lexical e baseada em aprendizado de máquina (AM). Por fim, esta seção é concluída com a apresentação de trabalhos correlatos.

A. Abordagem Lexical

As abordagens lexicais usam como base de conhecimento recursos denominados *léxicos de opinião* [13], que associam palavras à sua orientação de sentimento (por exemplo, *scores* positivos e negativos). Seu uso em AS parte do pressuposto de que palavras isoladas podem ser consideradas como uma unidade de informação de opinião e, portanto, podem fornecer indicações para detectar o sentimento e a subjetividade do documento [14]. Para definição da classificação de polaridade de palavras em um texto, normalmente recorre-se a dicionários léxicos.

Os dicionários léxicos são um conjunto de palavras previamente polarizadas que visam auxiliar na classificação de um texto. Dentre os principais dicionários disponíveis, pode-se mencionar o *Harvard Dictionary* e o *SentiWordNet* [15], voltados para a língua inglesa, e o *SentiLex-PT/BR* [16], voltado para o português.

O *SentiLex-PT/BR* é o dicionário de propósito geral concebido para AS e opinião pioneiro para a língua portuguesa, sendo atualmente constituído por 7.014 lemas e 82.347 formas flexionadas. As palavras apresentam polaridades positiva, negativa e neutra, valoradas respectivamente como +1, -1 e 0 [17].

Por outro lado, o dicionário proposto por Tim Loughran e Bill McDonald corresponde a uma adaptação do *Harvard Dictionary* contendo uma lista de palavras voltada a textos financeiros [18]. Esse dicionário não possui versão em português até o momento, sendo uma das tarefas deste trabalho propor a sua adaptação a essa língua.

Apesar do uso de dicionários léxicos ser simples e intuitivo, métodos léxicos baseados em dicionários genéricos podem apresentar problemas em domínios específicos, uma vez que certas palavras podem ter significados e, consequentemente, polaridades distintas em domínios diferentes. Loughran & McDonald [18] apontaram que, para o domínio do mercado financeiro, quase 75% das palavras identificadas como de polaridade negativa pelo *Harvard Dictionary* são tipicamente não-negativas em textos de temática financeira.

Há diversos métodos utilizados para classificar um texto a partir das notas de polaridade das palavras contidas em um dicionário léxico. Em [19] é proposta a soma das pontuações dos termos (*term score summation*) e em [20] a posição dos adjetivos (*adjectives position*), em que há a busca por adjetivos imediatamente antes e depois das palavras do dicionário, e o método base (*baseline*), em que há a análise de adjetivos três posições antes e depois das palavras do dicionário.

B. Abordagem baseada em Aprendizado de Máquina

As técnicas de Aprendizado de Máquina para AS têm como objetivo treinar um classificador, utilizando documentos associados a rótulos conhecidos (por ex., positivo ou negativo), para posterior classificação de documentos cujos rótulos são desconhecidos. De acordo com [21] e [22], técnicas clássicas como o algoritmo *Naive Bayes* têm apresentado boa acurácia na classicação de textos tradicionais, além de estarem entre as mais comuns na literatura da área.

O *Naive Bayes* é um classificador estatístico, baseado no Teorema de Bayes, que é capaz de prever a probabilidade de uma amostra pertencer a uma determinada classe. O classificador é denominado "ingênuo" (*naive*) por assumir que os atributos são condicionalmente independentes, ou seja, as informações sobre um evento não estão correlacionadas umas com as outras. Mesmo que esses atributos sejam dependentes uns dos outros, para o classificador *Naive Bayes* todas estas propriedades contribuem de forma independente. Essa suposição é responsável por simplificar os cálculos computacionais [23].

Por outro lado, nos últimos anos a literatura tem voltado a sua atenção para as redes neurais artificiais (RNA), que são inspiradas nos modelos de processamento sensorial feito pelo cérebro e podem ser criadas simulando uma rede de modelos de neurônios em um computador. Aplicando algoritmos que imitam os processos de neurônios reais, uma RNA pode ser capaz de aprender a resolver diferentes tipos de problemas [24]. A perceptron multicamadas (MLP - do inglês, *Multilayer Perceptron*) é uma rede neural artificial utilizada para modelar funções não lineares [25]. Ela baseia-se em uma camada de entrada que recebe os dados fornecidos à rede, para que possam ser ponderados pelos pesos sinápticos das conexões entre os neurônios presentes nas n camadas ocultas. Por fim, a função de ativação atua sobre o resultado anterior, que é transferido à camada de saída, composta por p neurônios, sendo que p é o número de classes envolvidas no problema [4] [25].

Esse trabalho, portanto, adota tanto o *Naive Bayes* como a MLP como técnicas de AM voltadas para a tarefa de Análise de Sentimentos.

C. Trabalhos Correlatos

Diversos trabalhos na literatura apresentam o uso de técnica lexical para AS em domínios específicos. Os autores do trabalho [26] mediram o sentimento em publicações no Twitter para prever o mercado financeiro com duas ferramentas lexicais de monitoramento de sentimentos e obtiveram precisão de 86,7% na previsão do movimento dos valores de fechamento do índice *Dow Jones Industrial Average* (DJIA). Os autores do trabalho [9] extraíram e analisaram os efeitos dos sentimentos de notícias sobre ações de empresas do setor farmacêutico, apresentando resultados que atingiram 70,59% de acerto na previsão das tendências de movimento dos preços das ações no curto prazo. Já os autores de [10] aplicaram uma abordagem léxica para analisar o sentimento de notícias relacionadas a ações individuais listadas na Bolsa de Valores de Kuala

Lumpur (KLSE) usando o método Léxico, registrando de 74,0% a 79,1% de acurácia.

Por outro lado, a literatura também apresenta resultados interessantes de AS com o uso de Aprendizado de Máquina. Em [27], há um estudo comparativo entre técnicas de AM e abordagem lexical, com o uso do dicionário *SentiWordNet*, para rastrear os sentimentos expressos em dados publicados *online* sobre produtos. Como resultado, o *Naive Bayes* e a MLP mostraram-se superiores às outras técnicas estudadas. Os autores do trabalho [28] implementaram cinco modelos de AS baseados em Aprendizado de Máquina e avaliaram seus desempenhos na previsão das tendências do mercado de ações. Os resultados experimentais mostram que a *Random Forest* levou a melhores resultados para grandes conjuntos de dados, enquanto o *Naive Bayes* foi o melhor para conjuntos de dados pequenos. Em [29], foi proposto um sistema de classificação de textos da área da saúde e sua avaliação de desempenho comparando diversas técnicas. A rede neural (MLP) obteve o melhor desempenho dentre os cinco tipos de classificadores estudados, atingindo 92% de acurácia.

É importante destacar, finalmente, que existem poucos trabalhos correlatos que desenvolvem AS voltada à língua portuguesa. Em [30], a AS em língua portuguesa é realizada por meio de junção de diferentes léxicos adaptados ao idioma português, enquanto que [31] apresenta uma estratégia com léxicos para AS em língua portuguesa usando o conhecimento linguístico sobre o contexto ao qual a palavra está incluída na sentença. O presente trabalho contribui para apresentar um comparativo de técnicas de AS para o português do Brasil, buscando complementar os trabalhos que adotam apenas analisadores léxicos.

Com relação aos trabalhos da literatura e ao panorama apresentado, o presente trabalho apresenta os seguintes diferenciais: (1) o desenvolvimento de dois dicionários financeiros adaptados do inglês para a língua portuguesa, capazes de permitir a melhora de métodos léxicos aplicados a essa língua; (2) uma estratégia de combinações de diferentes dicionários para AS de textos financeiros em português; e (3) a comparação entre técnicas de Aprendizado de Máquina e o dicionário financeiro proposto. Por fim, este trabalho também contribui para a literatura no sentido de que, de acordo com o constatado por [4], existem poucos trabalhos em língua portuguesa voltados à AS aplicada ao mercado financeiro.

III. Metodologia

Nesta seção, abordaremos a metodologia adotada neste trabalho, desde o processo de coleta dos dados e seu pré-processamento até as técnicas utilizadas para a AS. Todos os códigos e dados produzidos nesse trabalho estão disponíveis no Repositório de Dados de Pesquisa da Universidade Estadual de Campinas (REDU) [32].

A. Obtenção dos Dados

As empresas escolhidas para este estudo foram: Bradesco, Petrobras, Vale, Magazine Luíza e Gol, uma vez que as ações destas instituições são constantemente negociadas, deixando-as como ativas no mercado de ações brasileiro. Em virtude disso, também é comum encontrar notícias publicadas *online* sobre estas empresas. Ao todo foram coletadas 828 notícias, dentro de um período de aproximadamente 3 anos, extraídas dos sites Folha de São Paulo, Estadão, InfoMoney, MoneyTimes, Portal Exame, Último Segundo, UOL, Yahoo, G1 e Estadão com o auxílio de um *script* desenvolvido em Python.

A partir das notícias coletadas, foram rotuladas manualmente 555 com o rótulo positivo e 273 com o negativo, considerando a seguinte perspectiva: o rótulo positivo significa notícia favorável sobre uma determinada ação ao investidor (i.e., expectativa de alta) e rótulo negativo significa uma notícia desfavorável ao investidor (i.e., expectativa de baixa).

B. Pré-processamento das Bases de Dados e Dicionários

Para aplicar as técnicas de classificação textual, é imprescindível realizar o pré-processamento da base de dados a fim de modelar o texto para que a máquina entenda. Para a fase de pré-processamento, as seguintes etapas foram realizadas, conforme [4], tanto nas bases de dados quanto nos dicionários: *tokenization* (extração de unidades mínimas de texto a partir de um texto livre), normalização (conversão de caracteres maiúsculos para minúsculos e a remoção de caracteres especiais e números), remoção de *stopwords* (termos não representativos para um documento) e *stemming* (reduzir todas as possíveis variações de uma palavra a uma forma única: sua raiz ou radical). Todas estas fases foram efetuadas com o auxílio das bibliotecas *RE* e *NLTK* disponibilizadas para Python.

C. Abordagem Lexical

Para aplicar a abordagem léxica, realizamos a adaptação do *Loughran-McDonald Dictionary* para o português com o auxílio da ferramenta *online Google Translate* e ajustamos sua polaridade. Outrossim, conforme apresentado na Seção de Resultados, o uso dos dicionários lexicais atingiram apenas cerca de 50% de acurácia, e, a fim de os aperfeiçoar, combinamos o dicionário *SentiLex-PT/BR* ao dicionário financeiro traduzido. Por fim, através dos dicionários lexicais, utilizamos um analisador léxico que emprega a técnica de soma das pontuações dos termos, em conformidade com [19], para obtermos como saída o sentimento resultante do documento. Para investigar os resultados obtidos, o analisador léxico calcula a acurácia da AS [33], bem como também a métrica *F1-Score*. Enquanto a acurácia mede a quantidade de elementos corretamente classificados divididos pelo total de elementos da base de dados, a métrica *F1-Score* considera a média harmônica entre as métricas precisão e revocação, que medem a relação entre instâncias classificadas e a classe a que pertencem, conforme detalhado em [4].

A seguir, são apresentados os detalhes de como foram executadas estas tarefas.

1) Ajuste de polaridade do Loughran-McDonald Dictionary: Para que fosse possível realizar uma comparação adequada utilizando o dicionário financeiro adaptado e o *SentiLex-PT/BR*, foi necessário converter as categorias das palavras em pesos numéricos, o que levou a duas versões deste dicionário.

Na Versão I, as palavras classificadas como *negative*, *uncertainly*, *litigious* e *constraining* receberam pesos iguais a

-1 por apresentarem teor negativo, enquanto as demais com teor positivo, *strong modal*, *weak modal* e *positive*, receberam peso +1. Na Versão II, apenas mantivemos as palavras categorizadas como *negative*, que assumiram peso -1, e *positive*, que assumiram peso +1. A quantidade de palavras, total e por classe, de cada versão do dicionário financeiro e do SentiLex-PT/BR após os ajustes de polaridade e pré-processamento é dada na Tabela I. Pode-se observar que a quantidade de termos negativos é significativamente maior do que a de positivos para todos os dicionários abordados neste estudo. Vale a pena destacar que consideramos nos experimentos tanto dicionários com *stemming* quanto dicionários sem *stemming*.

2) Combinação dos dicionários SentiLex-PT/BR e Loughran-McDonald: Embora o dicionário financeiro proposto por Tim Loughran e Bill McDonald seja especializado em análise de mercado financeiro, ele foi criado manualmente. Portanto, as palavras de polaridade no dicionário eram apenas uma parte das palavras que apareciam em dicionários utilizados como base [34], assim resultando em um dicionário com poucas palavras polarizadas. Por conseguinte, a presença de palavras que exprimem opinião em um texto, mas que não apresentam polaridade definida no dicionário, podem afetar o desempenho do classificador de sentimentos. Como solução, propusemos neste trabalho a combinação do dicionário *SentiLex-PT/BR*, que apresenta uma quantidade de palavras significativamente maior, com as duas versões do dicionário financeiro traduzido, assim, gerando dois dicionários combinados.

Para a combinação, o analisador léxico primeiro verifica se a palavra está contida no dicionário financeiro, caso não esteja, ele verificará se está contida no *SentiLex-PT/BR*. Assim, o algoritmo prioriza a polaridade da palavra fornecida pelo dicionário financeiro.

D. Abordagem de Aprendizado de Máquina

A fim de realizar a AS utilizando algoritmos de Aprendizado de Máquina, os algoritmos utilizados foram: *Naive Bayes* e a RNA do tipo MLP, ambos implementados na linguagem Python com o auxílio da biblioteca *scikit-learn*. A seguir, são detalhados os parâmetros e entradas de cada um desses algoritmos.

1) Naive Bayes: Para realizar a classificação supervisionada, foi utilizado o modelo *MultinomialNB* disponível para Python. Foi considerada como entrada do algoritmo uma frequência mínima de termos e a representação dos nossos dados adotando os modelos: *bag-of-words* (BOW), *term-frequency* (TF) e *term frequency-inverse document frequency* (TF-IDF) [4]. Enquanto a matriz denominada *bag-of-words* corresponde à união de todos os vetores *one-hot*, o cálculo da frequência de cada termo (*term-frequency*) corresponde à frequência em que um termo aparece no documento. Por fim, a frequência inversa do documento (*term frequency-inverse document frequency*) multiplica a frequência com que um termo ocorre num documento pelo inverso da frequência nos documentos [35].

Nos experimentos, utilizamos as mesmas bases de dados usadas na aplicação da abordagem lexical, contudo, devido ao seu tamanho relativamente pequeno, optamos pela técnica de validação cruzada por meio do método *k-fold*, com k=10. Esse método consiste em dividir aleatoriamente o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disso, um subconjunto é utilizado como conjunto de teste e os *k-1* restantes são utilizados para treinar o modelo. Por fim, a medida de desempenho é dada pela média dos k valores calculados junto às partições de teste, que não foram utilizadas durante a etapa de treinamento.

2) Rede Neural Perceptron Multicamadas: Para a execução da rede, foram definidos parâmetros conforme apresentado a seguir. As entradas são definidas como vetores representando palavras por meio das abordagens: BOW, TF e TF-IDF, conforme também utilizado no algoritmo *Naive Bayes*. Além disso, foram utilizadas nas representações as 1500 palavras mais frequentes, sendo incluídas apenas palavras que ocorrem em pelo menos 3 documentos e palavras que ocorram em no máximo 70% de todos os documentos.

A *MLP* utilizada nos experimentos foi ajustada para treinamento com no máximo 1000 épocas. Para testar a rede neural também foi adotada a técnica de validação cruzada *k-fold*, sendo k=10, e para avaliá-la, usamos acurácia e *F1-Score*.

IV. Resultados Experimentais

Nesta seção, serão apresentados os resultados experimentais obtidos neste trabalho. Os resultados do emprego da abordagem lexical são discutidos na Seção IV-A, os da aplicação do *Naive Bayes* na Seção IV-B e da MLP na Seção IV-C.

A. Abordagem Lexical

Foram utilizadas como entrada para o analisador lexical duas bases de dados distintas: a primeira composta pelo conteúdo textual das notícias e a segunda resultante da aplicação de *stemming* na primeira. Para cada base de dados, aplicamos os seguintes dicionários e suas respectivas versões *stemming*:

1) Dicionário Financeiro Versão I;
2) Dicionário Financeiro Versão II;
3) *SentiLex-PT/BR*;
4) Combinação do Dicionário Financeiro Versão I com o *SentiLex-PT/BR*;
5) Combinação do Dicionário Financeiro Versão II com o *SentiLex-PT/BR*;

A média das acurácias e de *F1-Score*, em porcentagem, obtidas para as duas bases de dados são exibidas na Tabela II. Pode-se observar que as acurácias resultantes do emprego do Dicionário Financeiro Versão II combinado com o SentiLex-PT/BR foram as maiores, em especial os 58,2% de acurácia oriundos da aplicação deste dicionário na base de notícias sem a aplicação de *stemming*, que também obteve o valor mais significativo de *F1-Score*: 58,8%.

Em detalhes, o algoritmo classificou corretamente 52% das notícias positivas e 85% das negativas. Isso pode ter ocorrido em razão dos termos dos dicionários serem majoritariamente negativos, como visto na Tabela I, de modo que a probabilidade de conter palavras com pesos negativos em um texto seja maior, e consequentemente, fazendo com que a soma das pontuações dos termos resultante também seja negativa. Essa

TABELA I
QUANTIDADE DE PALAVRAS DOS DICIONÁRIOS, CONSIDERANDO VERSÕES COM *stemming* E SEM *stemming*.

	SentiLex PT/BR	SentiLex PT/BR Stemming	Dic. Financeiro Versão I	Dic. Financeiro Versão I Stemming	Dic. Financeiro Versão II	Dic. Financeiro Versão II Stemming
Positivo	3939	2830	341	213	328	197
Negativo	6561	4975	2913	1628	1940	1056
Neutro	35768	29876	0	0	0	0
Total	46268	37681	3254	1841	2268	1253

TABELA II
MÉDIA DAS ACURÁCIAS E F1-SCORE (EM %) OBTIDAS COM A APLICAÇÃO DO MÉTODO LÉXICO.

	Dic. Financeiro Ver. I		Dic. Financeiro Ver. II		SentiLex PT/BR		Dic. Combinado Ver. I		Dic. Combinado Ver. II	
	Acurácia	**F1-score**	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**
Stemming	38,4	35,0	54,3	56,8	47,5	48,3	38	33,8	57,1	57,4
Sem Stemming	38,7	37,3	46,8	0,5	52,7	53,5	42,7	40,8	58,2	58,8

técnica classifica com maior acurácia notícias negativas em detrimento das positivas, o que influencia desfavoravelmente na acurácia final dos experimentos realizados, uma vez que a quantidade de notícias positivas é mais do que o dobro das negativas na base de dados utilizada.

B. Naive Bayes

Para cada base de dados, realizamos uma série de execuções distintas do *Naive Bayes* em Python, alterando a frequência mínima de termos e o modelo de representação de dados (BOW, TF e TF-IDF) e utilizando k=10 para a técnica *K-fold* de validação cruzada. As médias das acurácias e *F1-Score*, obtidas em cada experimento, são exibidas nas Tabelas III e IV, respectivamente, considerando os resultados com *stemming* e sem *stemming*.

Obtivemos resultados que mostram a capacidade do *Naive Bayes* para classificar corretamente sentimentos em textos. Os resultados obtidos neste experimento foram superiores aos obtidos pela abordagem lexical, sendo a maior acurácia de 75,36% obtida para a base de dados sem a aplicação de *stemming* utilizando as seguintes configurações: frequência mínima de termos = 10,0 e uso de TF-IDF. Analisando a métrica *F1-score* resultante deste experimento, podemos observar que o algoritmo apresentou como melhor valor 84,02%, para a mesma configuração apresentada.

C. Rede Neural Perceptron Multicamadas

Para cada base de dados, a rede neural do tipo MLP foi configurada com os seguintes valores: (i) 1, 3, 5 e 10 camadas ocultas e (ii) 10, 25, 50 e 100 neurônios por camada oculta. As médias das acurácias obtidas para as duas bases de dados – exibidas nas Tabelas V e VI, respectivamente, com *stemming* e sem *steming*.

Utilizando como entrada da rede a abordagem BOW, obtivemos 78,98% de acurácia, o maior resultado dentre todos os experimentos realizados. Este valor foi atingido utilizando a base de dados com a aplicação de *stemming* e a RNA configurada com 10 camadas ocultas e 25 neurônios em cada camada. Considerando a métrica F1-Score, observamos que o algoritmo apresentou melhor resultado com valor 85,01% com a configuração que utiliza BOW e *stemming* como entradas da rede, 3 camadas e 50 neurônios.

Em geral, vale a pena observar que, para a MLP, os experimentos realizados na base de dados com *stemming* obtiveram melhores médias de acurácia do que os realizados na base sem *stemming*. As outras abordagens estudadas apresentaram melhor desempenho utilizando a base de dados sem a aplicação de *stemming*.

V. CONCLUSÃO

Este trabalho apresentou desenvolvimento de análise de sentimentos aplicado ao domínio financeiro no idioma português brasileiro. Como frutos deste trabalho, apresentamos: (*i*) uma base de dados composta por 828 notícias rotuladas manualmente; (*ii*) dois dicionários lexicais voltados para o domínio financeiro e adaptados para o português, bem como a proposta de uma estratégia de uso combinado de diferentes dicionários léxicos para a área financeira em português; e (*iii*) a realização de um estudo comparativo entre três estratégias diferentes de análise automática de sentimentos: a abordagem lexical, o algoritmo de aprendizagem de máquina *Naive Bayes* e a rede neural perceptron multicamadas.

Com os resultados obtidos, verificamos que o algoritmo *Naive Bayes* e a rede MLP obtiveram acurácias e F1-scores maiores na AS em relação à abordagem lexical, assim como nos trabalhos propostos por [27] e [36]. Apesar da abordagem lexical não apresentar resultados tão bons quanto os obtidos pela aplicação das técnicas de Aprendizado de Máquina, vale enfatizar que a melhor acurácia obtida para estas abordagens foi com o uso do dicionário combinado, ou seja, o léxico financeiro do *Loughran-McDonald Dictionary*, considerando apenas as classes *negative* e *positive*, adaptado do inglês para o português e combinado com o *SentiLex-PT/BR*. De fato, conforme apresentado em [31], apesar de existirem técnicas mais elaboradas para o uso de dicionários léxicos, dificilmente é possível atingir resultados de AS equivalentes às técnicas de AM em um domínio específico. Além disso, o trabalho voltado ao domínio financeiro em português [30] apresenta resultados que demonstram a necessidade da combinação de

TABELA III
MÉDIA E DESVIO PADRÃO DAS ACURÁCIAS E F1-SCORE (EM %). ALGORITMO *Naive Bayes* COM STEMMING.

	Bag of Words		TF		TF-IDF	
Freq.	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**
3	75,23 ± 0,04	81,43 ± 0,03	73,67 ± 0,03	83,42 ± 0,02	74,87 ± 0,04	83,86 ± 0,03
5	75,11 ± 0,04	81,32 ± 0,02	73,76 ± 0,03	83,40 ± 0,02	74,98 ± 0,04	83,86 ± 0,03
10	74,39 ± 0,04	80,86 ± 0,03	73,55 ± 0,03	83,43 ± 0,02	74,90 ± 0,04	83,83 ± 0,03

TABELA IV
MÉDIA E DESVIO PADRÃO DAS ACURÁCIAS E F1-SCORE (EM %). ALGORITMO *Naive Bayes* SEM STEMMING.

	Bag of Words		TF		TF-IDF	
Freq.	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**
3	74,16 ± 0,05	80,24 ± 0,04	73,42 ± 0,03	83,21 ± 0,01	75,00 ± 0,03	83,79 ± 0,02
5	74,28 ± 0,05	80,34 ± 0,04	73,42 ± 0,03	83,21 ± 0,01	74,87 ± 0,03	83,72 ± 0,02
10	73,31 ± 0,04	79,71 ± 0,03	73,30 ± 0,02	83,15 ± 0,01	75,36 ± 0,03	84,02 ± 0,01

TABELA V
MÉDIA E DESVIO PADRÃO DAS ACURÁCIAS E F1-SCORE. REDE MLP COM STEMMING.

		Bag of Words		TF		TF-IDF	
Camadas	**Neurônios**	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**
1	**10**	77,66 ± 0,04	84,03 ± 0,03	77,18 ± 0,05	83,45 ± 0,03	78,14 ± 0,04	84,25 ± 0,03
	25	78,02 ± 0,04	84,36 ± 0,03	77,06 ± 0,05	83,48 ± 0,03	77,41 ± 0,04	83,71 ± 0,03
	50	77,90 ± 0,04	84,19 ± 0,03	77,18 ± 0,04	83,50 ± 0,03	78,26 ± 0,04	84,33 ± 0,03
	100	77,90 ± 0,04	84,31 ± 0,03	77,42 ± 0,05	83,68 ± 0,04	77,66 ± 0,04	83,89 ± 0,03
3	**10**	77,66 ± 0,05	83,84 ± 0,03	77,42 ± 0,05	83,97 ± 0,03	78,26 ± 0,05	84,36 ± 0,03
	25	76,81 ± 0,04	83,41 ± 0,03	77,54 ± 0,05	83,77 ± 0,03	77,90 ± 0,04	84,12 ± 0,03
	50	78,87 ± 0,05	85,01 ± 0,03	77,18 ± 0,04	83,61 ± 0,03	77,30 ± 0,05	83,65 ± 0,03
	100	77,53 ± 0,04	84,16 ± 0,02	77,29 ± 0,05	83,64 ± 0,03	78,14 ± 0,05	84,33 ± 0,03
5	**10**	78,02 ± 0,04	84,3 ± 0,03	75,25 ± 0,06	82,15 ± 0,04	77,06 ± 0,05	83,61 ± 0,03
	25	75,37 ± 0,06	82,12 ± 0,04	77,90 ± 0,05	84,17 ± 0,03	77,78 ± 0,05	84,08 ± 0,04
	50	78,02 ± 0,05	84,24 ± 0,04	77,42 ± 0,04	83,89 ± 0,03	77,78 ± 0,05	84,10 ± 0,03
	100	76,33 ± 0,05	83,01 ± 0,04	77,17 ± 0,04	83,59 ± 0,03	77,29 ± 0,05	83,49 ± 0,03
10	**10**	77,30 ± 0,03	83,38 ± 0,02	76,33 ± 0,05	83,09 ± 0,03	76,96 ± 0,04	82,74 ± 0,04
	25	78,98 ± 0,04	84,71 ± 0,03	77,41 ± 0,05	83,88 ± 0,03	77,66 ± 0,04	84,04 ± 0,03
	50	75,96 ± 0,04	82,26 ± 0,03	78,14 ± 0,04	84,28 ± 0,03	77,77 ± 0,04	83,82 ± 0,03
	100	77,17 ± 0,06	83,46 ± 0,04	76,69 ± 0,05	83,04 ± 0,03	77,05 ± 0,04	83,22 ± 0,03

TABELA VI
MÉDIA E DESVIO PADRÃO DAS ACURÁCIAS (EM %) E F1-SCORE. REDE MLP SEM STEMMING.

		Bag of Words		TF		TF-IDF	
Camadas	**Neurônios**	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**	**Acurácia**	**F1-score**
1	**10**	75,61 ± 0,03	82,20 ± 0,02	75,84 ± 0,03	82,58 ± 0,02	75,36 ± 0,03	82,21 ± 0,02
	25	75,73 ± 0,04	82,41 ± 0,03	75,72 ± 0,03	82,46 ± 0,02	75,61 ± 0,04	82,26 ± 0,03
	50	76,57 ± 0,03	83,07 ± 0,02	75,97 ± 0,04	82,66 ± 0,03	75,73 ± 0,03	82,40 ± 0,02
	100	76,81 ± 0,03	83,41 ± 0,02	75,84 ± 0,04	82,56 ± 0,03	75,24 ± 0,04	82,08 ± 0,03
3	**10**	75,25 ± 0,04	82,14 ± 0,03	75,84 ± 0,03	82,58 ± 0,02	75,49 ± 0,03	82,31 ± 0,02
	25	75,85 ± 0,04	82,51 ± 0,03	76,69 ± 0,04	83,23 ± 0,03	75,36 ± 0,03	82,15 ± 0,02
	50	76,93 ± 0,04	83,38 ± 0,03	75,72 ± 0,04	82,59 ± 0,03	75,84 ± 0,03	82,61 ± 0,02
	100	77,53 ± 0,03	83,97 ± 0,03	76,45 ± 0,02	83,10 ± 0,02	76,09 ± 0,02	82,89 ± 0,01
5	**10**	75,00 ± 0,03	82,01 ± 0,02	76,09 ± 0,04	82,90 ± 0,03	76,33 ± 0,03	82,99 ± 0,03
	25	76,93 ± 0,03	83,38 ± 0,02	75,60 ± 0,03	82,47 ± 0,02	76,69 ± 0,03	83,29 ± 0,02
	50	77,90 ± 0,02	84,15 ± 0,02	77,05 ± 0,03	83,34 ± 0,02	77,18 ± 0,02	83,61 ± 0,01
	100	75,71 ± 0,03	82,66 ± 0,02	75,72 ± 0,03	82,31 ± 0,02	77,17 ± 0,04	83,48 ± 0,03
10	**10**	75,00 ± 0,04	81,66 ± 0,03	74,15 ± 0,04	80,32 ± 0,04	75,60 ± 0,04	82,50 ± 0,02
	25	74,28 ± 0,04	80,90 ± 0,03	76,45 ± 0,04	82,86 ± 0,03	75,36 ± 0,03	81,82 ± 0,03
	50	76,56 ± 0,03	83,06 ± 0,02	76,09 ± 0,03	82,82 ± 0,02	75,48 ± 0,04	81,77 ± 0,03
	100	76,81 ± 0,03	83,07 ± 0,02	75,84 ± 0,03	82,44 ± 0,03	75,48 ± 0,03	81,69 ± 0,03

dicionários para melhorar a AS com base em léxicos. Nesse ponto, o diferencial do nosso trabalho é apresentar a melhora nos resultados quando consideramos o uso de um dicionário de propósito específico da área financeira.

Como trabalhos futuros, pretendemos desenvolver uma estratégia de investimento baseado no resultado da AS desenvolvida neste trabalho.

AGRADECIMENTOS

Este estudo foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001, e pelo processo nº 2018/24371-1, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

REFERÊNCIAS

[1] https://valorinveste.globo.com/mercados/renda-variavel/noticia/2020/05/06/numero-de-investidores-pessoa-fisica-na-bolsa-sobe-a-238-milhoes-em abril.ghtml

[2] Z. Madhoushi, A. R. Hamdan, and S. Zainudin, "Sentiment analysis techniques in recent works," in *2015 Science and Information Conference (SAI)*, pp. 288–291, 2015.

[3] S. P. Kothari and B. J. Warner, *The econometrics of event studies*, vol. 1, pp. 3–36. Elsevier, 2006.

[4] A. Carosia, G. Coelho, and A. Silva, "The influence of tweets and news on the brazilian stock market through sentiment analysis," pp. 385–392, 10 2019.

[5] B. Liu, "Sentiment analysis and subjectivity," in *Handbook of Natural Language Processing*, Taylor and Francis, 2 ed., 2010.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comp. Linguistics*, vol. 37, pp. 267–307, 2011.

[7] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proc. of the 40th Meeting of the Association for Computational Linguistics*, p. 417–424, 2002.

[8] E. Alpaydin, *Introduction to Machine Learning*. MIT Press; fourth edition, 4 ed., 2020.

[9] D. Shah, H. Isah, and F. Zulkernine, "Predicting the effects of news sentiments on the stock market," in *Proc. of the IEEE International Conference on Big Data*, pp. 4705–4708, 2018.

[10] T. L. Im, P. W. San, C. K. On, R. Alfred, and P. Anthony, "Analysing market sentiment in financial news using lexical approach," in *Proc. of the IEEE Conference on Open Systems (ICOS)*, pp. 145–149, 2013.

[11] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proc of the 2nd International Conference on Knowledge Capture*, pp. 70–77, 2003.

[12] F. Benevenuto, F. Ribeiro, and M. Araújo, "Métodos para análise de sentimentos em mídias sociais," in *Proc. of the Brazilian Symposium on Multimedia and the Web (Webmedia)*, 2015.

[13] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc of the ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, p. 168–177, 2004.

[14] F. Chiavetta, B. Giosué, and G. Pilato, "A lexicon-based approach for sentiment classification of amazon books reviews in italian language," in *Proc. of the 2nd. Intl. Conference on Web Information Systems and Technologies*, 2016.

[15] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proc. of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pp. 417–422, 2006.

[16] M. J. Silva, P. Carvalho, P. Costa, and L. Sarmento, "Automatic expansion of a social judgment lexicon for sentiment analysis," Tech. Rep. TR 10-08, University of Lisbon, 2010.

[17] P. Carvalho and M. J. Silva, "SentiLex-PT: Principais características e potencialidades," *Linguística, Informática e Tradução: Mundos que se Cruzam*, vol. 7, pp. 425–439, 2015.

[18] T. Loughran and B. Mcdonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *Journal of Finance, Forthcoming*, vol. 66, pp. 35–65, 2010.

[19] A. Hamouda and M. Rohaim, "Reviews classification using sentiwordnet lexicon," in *Proc. of the World Congress on Computer Science and Information Technology*, vol. 2, pp. 2090–4517, 2011.

[20] L. D. Freitas and R. Vieira, "Exploring resources for sentiment analysis in portuguese language," in *Proc. of the 2015 Brazilian Conference on Intelligent Systems (BRACIS)*, p. 152–156, 2015.

[21] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. of the Conference on Empirical Methods in NLP*, p. 79–86, 2002.

[22] B. Pang, L. Lee, *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, pp. 1–135, 2008.

[23] G. Babin, K. Stanoevska-Slabeva, and P. Kropf, "E-technologies: Transformation in a connected world," in *Proc. of the 5th International Conference on on E-Technologies (MCETECH)*, Springer-Verlag Berlin Heidelberg, 2011.

[24] A. Krogh, "What are artificial neural networks?," *Nature Biotechnology*, vol. 26, p. 195–197, 2008.

[25] M. Bounabi, K. El Moutaouakil, and K. Satori, "A probabilistic vector representation and neural network for text classification," in *Big Data, Cloud and Applications* (Y. Tabii, M. Lazaar, M. Al Achhab, and N. Enneya, eds.), (Cham), pp. 343–355, Springer International Publishing, 2018.

[26] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, pp. 1–8, 2011.

[27] S. Ahmed and A. Danti, "A novel approach for sentimental analysis and opinion mining based on sentiwordnet using web data," in *2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15)*, pp. 1–5, 2015.

[28] I. Kumar, K. Dogra, C. Utreja, and P. Yadav, "A comparative study of supervised machine learning algorithms for stock market trend prediction," in *Proc. of the 2nd International Conference on Inventive Communication and Computational Technologies*, pp. 1003–1007, 2018.

[29] S. K. Srivastava, S. K. Singh, and eJasjit S. Suri, "Healthcare Text Classification System and its Performance Evaluation: A Source of Better Intelligence by Characterizing Healthcare," *Journal of Medical Systems*, no. 97, 2018.

[30] R. F. Martins, A. Pereira, and F. Benevenuto, "An approach to sentiment analysis of web applications in portuguese," *Proceedings of the 21st Brazilian Symposium on Multimedia and the Web*, 2015.

[31] L. V. Avanço and M. d. G. V. Nunes, "Lexicon-based sentiment analysis for reviews of products in brazilian portuguese," *2014 Brazilian Conference on Intelligent Systems*, 2014.

[32] B. Januário, A. Carosia, G. Coelho, and A. Silva, "Financial news about brazilian companies listed on b3 and source-codes to perform sentiment analysis," *Repositório de Dados de Pesquisa da Unicamp*, 2021.

[33] L. N. de Castro and D. G. Ferrari, *Introdução a mineração de dados*. SARAIVA, 1 ed., 2017.

[34] R. To, K. Izumi, H. Sakaji, and S. Suda, "Lexicon creation for financial, sentiment analysis using network embedding," *Journal of Mathematical Finance*, pp. 896–907, 2017.

[35] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[36] Z. Hailong, G. Wenyan, and J. Bo, "Machine learning and lexicon based methods for sentiment classification: A survey," in *Proc. of the 11th Web Information System and Application Conference*, pp. 262–265, 2014.

Brenda Alexsandra Januário Brenda Alexsandra Januário is graduating in Information Systems at the School of Technology of the University of Campinas - UNICAMP. She is currently a Data Engineer at Itaú Unibanco and her areas of interest are cloud computing, data mining and machine learning.

Arthur Emanuel de Oliveira Carosia is a Ph.D. student at the School of Technology of the University of Campinas, Brazil. He is a Professor at the Instituto Federal de São Paulo (IFSP), Brazil. His research interests are Data Science and Computational Intelligence, with recent work on machine learning and stock market prediction.

Ana Estela Antunes da Silva has an undergraduation degree in Computer Science from the University of Campinas – UNICAMP, a Master degree in Computer Science from Massey University and a Ph.D. in Computer Engineering from the University of Campinas. She is currently a teacher in the School of Technology at University of Campinas. Has experience in the area of Computer Science, with emphasis on: text mining, data mining and intelligent systems for decision making.

Guilherme Palermo Coelho is a Computer Engineer (University of Campinas - UNICAMP), with M.Sc. and Ph.D. degrees in Electrical Engineering (also from UNICAMP). He is an IEEE Senior Member and currently an Assistant Professor at the School of Technology of the University of Campinas, Brazil. His research interests are associated with Computational Intelligence in general, with recent work on metaheuristics for optimization (single and multi-objective), data mining, and machine learning.