Towards the Categorization of Brazilian Financial Market Headlines

Matheus Schmitz, Roger Immich, Gustavo Pessin and Geraldo P. Rocha Filho

***Abstract*—Financial market news portals are valuable sources of information as they hold great power over investors' decision-making processes. Due to the vast amount of text data produced by news portals, several studies have been conducted to comprehend the behavioral variations of texts and automate the categorization of short texts. However, extracting useful information that influences investors' decision-making process is not a trivial task, given that news portals use a heterogeneous and specific language for each content produced, making it challenging to generate a standard document format. This work proposes GOOSE, a solution for the cateGOrizatiOn of Short texts derived from multiple sources of information, to portray the financial market's current situation. To this end, GOOSE is based on Bidirectional Long Short-Term Memory (Bi-LSTM) and GloVe Embeddings to increase reliability in the short texts classification process. That way, GOOSE obtains data from news portals, which, once combined with a word embedding mechanism, are used as input for the Bi-LSTM to classify financial market news texts. The results obtained showed that GOOSE's efficiency in categorizing texts had an accuracy of 84% but also demonstrated the feasibility of its use in the extraction of information from financial market news portals.**

***Index Terms*—Financial Market, Machine Learning, Text Categorization**

I. Introdução

As bolsas de valores desempenham importante papel na economia mundial, sendo responsáveis pela negociação de títulos públicos, ações de empresas, contratos de *commodities*, dentre outros. A bolsa brasileira [B]^3 (Brasil, Bolsa, Balcão) foi oficialmente estabelecida em seu formato atual a partir de 2017, após a realização de fusões entre outras bolsas nacionais existentes. O site oficial [1] apresenta que o número de CPFs cadastrados quase quadruplicou em comparação com o registro do ano de 2018, em que passou de 813.291 para 3.229.318 registros ao final de 2020. Embora seja um número de investidores pequeno em relação à população do país, a tendência é de que esse quantitativo aumente.

Nesse contexto, os algoritmos de Aprendizado de Máquina (ML, do inglês *Machine Learning*) combinados com os avanços conquistados na área de Processamento de Linguagem Natural (NLP, do inglês *Natural Language Processing*) apresentam soluções para resolver o problema de categorização de textos curtos derivados de múltiplas fontes de informações no mercado financeiro [2], [3]. A categorização de textos consiste no processo de classificar os conteúdos textuais de acordo com categorias temáticas predefinidas, podendo ser aplicada para compreender as variações comportamentais do texto [3], [4]. Diversos desafios devem ser superados para obter um conhecimento útil no processo de classificação. A heterogeneidade linguística textual provida nos diversos portais de notícias e a variedade de conteúdo é um dos desafios que se destaca. Em razão disso, técnicas baseadas em regras são desafiadoras e podem não realizar de maneira eficiente o processo de categorização textual [5].

Diferentes trabalhos foram analisados para análise textual e extração de informação relacionada ao mercado financeiro [6]–[10]. Alguns dos trabalhos exploram a classificação binária do sentimento associado às publicações no que diz respeito ao mercado de ações [6]–[8]. Outros trabalhos propõem arquiteturas de rede neural ao combinar redes convolucionais e *embeddings* a nível de caracteres e redes recorrentes com Unidades Recorrentes Bloqueadas (GRU, do inglês *Gated Recurrent Units*) [11], [12]. Salienta-se, entretanto, que a modelagem de tais arquiteturas é complexa e, no geral, só atingem melhores resultados em comparação às arquiteturas mais simples com grande quantidade de dados rotulados.

Este trabalho propõe o GOOSE, uma solução para cateGOrização de textos curtOS derivados de múltiplas fontEs de informação para retratar a situação atual do mercado financeiro. Para tanto, o GOOSE é baseado em uma Bi-LSTM e *GloVe Embeddings* para aumentar a confiabilidade no processo de classificação dos textos curtos. Com isso, o GOOSE obtém os dados e seus respectivos rótulos dos portais de notícias que, quando combinados com um mecanismo de incorporação de palavras, são usados como entrada para a rede Bi-LSTM realizar a classificação dos textos de notícias do mercado financeiro. Em razão disso, as principais contribuições desta pesquisa são:

- Desenvolvimento de *crawlers* para a coleta de dados de múltiplas fontes de informação derivadas dos portais de notícias do mercado financeiro brasileiro;
- Execução e análise de um mecanismo de *Word Embeddings* para mapear palavras semanticamente semelhantes;
- Implementação e avaliação de uma Bi-LSTM modelada para a classificação de textos curtos relacionados ao mercado financeiro e que se adapta para novos conteúdos;

O restante deste trabalho é estruturado da seguinte maneira. A Seção II apresenta os trabalhos relacionados. A Seção III apresenta como foi desenvolvido o GOOSE, enquanto a sua validação é avaliada na Seção IV. Por fim, a Seção V apresenta

Matheus Schmitz, Geraldo P. Rocha Filho, Departamento de Ciência da Computação (CiC), Universidade de Brasília (UnB), e-mail: matheus.schmitz@aluno.unb.br, geraldof@unb.br.

Roger Immich, Instituto Metropole Digital (IMD), Universidade Federal do Rio Grande do Norte (UFRN), e-mail: roger@imd.ufrn.br

Gustavo Pessin, Instituto Tecnológico VAle (ITV), e-mail: gustavo.pessin@itv.org

as principais conclusões e os trabalhos futuros.

II. Trabalhos Relacionados

Esta seção apresenta trabalhos relacionados ao processo de categorização de textos do mercado financeiro. As pesquisas envolvendo dados de redes sociais e análise de sentimentos apresentam maior destaque na literatura, evidenciando a importância do processo de extração de informação com a classificação de dados textuais.

A combinação de diferentes tarefas de classificação em NLP foi utilizada para analisar a movimentação do índice *Dow Jones* em [6]. Os autores combinaram atividades como extração de tópicos, análise de sentimentos e reconhecimento de entidades nomeadas em textos extraídos do jornal *New York Times* e do *Twitter*. Os resultados indicam que as notícias exercem impacto em uma janela futura de dias. No entanto, o processo é dependente de ferramenta para a classificação prévia dos sentimentos e das entidades nomeadas presentes nos dados textuais. Este fato pode prejudicar o desempenho do modelo proposto, visto que as ferramentas não foram treinadas somente com textos relacionados ao mercado financeiro.

Os dados textuais em português de notícias e publicações no Twitter foram explorados com o objetivo de analisar a movimentação do índice Bovespa em [7]. A pesquisa relata a escassez de trabalhos envolvendo análise de sentimentos e a relação com o mercado de ações brasileiro. A metologia escolhida analisa o sentimento das notícias relacionadas ao mercado no dia anterior à abertura e compara o resultado com a movimentação registrada no índice no dia posterior. O modelo de classificação utiliza uma rede neural, que foi treinada a partir de notícias publicadas entre os anos de 2012 e 2014. Os rótulos foram atribuídos por especialistas, indicando a interferência do conteúdo na percepção do investidor. Entretanto, novas empresas são inseridas na bolsa de valores e novos assuntos são capazes de influenciar a tomada de decisão, sendo imprescindível a adição constante de conteúdos mais atualizados sobre as entidades financeiras.

Recentemente, o estudo desenvolvido em [8] explora a predição da movimentação de preços na bolsa de valores de Gana (GSE) ao combinar diferentes fontes de informação. O estudo combina dados textuais extraídos do Twitter, de fóruns e portais de notícias, além de métricas provenientes do *Google Trends*. Os resultados evidenciam que a combinação das fontes de dados aumenta a acurácia do modelo proposto. Também foi observado que uma janela com um maior número de dias é recomendada para agregar o sentimento geral, pois o resultado pode continuar influenciando o mercado em dias futuros. Apesar disso, a proposta utiliza poucos dados no treinamento da rede e depende de ferramenta para a classificação dos sentimentos nos textos, diferente desta pesquisa.

A análise de sentimentos em publicações do Twitter também foi explorada em [9]. Os autores utilizaram métodos estatísticos para identificar a relação de causalidade entre um índice de felicidade nas publicações da rede social [13] e o índice VIX de diferentes países. No entanto, o índice de felicidade contém eventos não relacionados ao mercado de ações, como feriados nacionais e internacionais. Adicionalmente, é avaliado com publicações em inglês, o que pode deixar de destacar eventos ocorridos em outras nacionalidades. A proposta utiliza dados rotulados, mas não apresenta um mecanismo de extração automática e não explora o conteúdo textual de maneira direta.

O impacto na movimentação dos índices *Standard and Poor's 500* (S&P 500 ou SPX) e Dow Jones (DJIA, do inglês *Dow Jones Industrial Average*) foi avaliado a partir das publicações no Twitter do ex-presidente americano Donald Trump [10]. Os autores utilizaram a ferramenta VADER [14] para classificar de maneira automática o sentimento dos *tweets* publicados durante o horário de funcionamento da bolsa americana (NYSE, do inglês *New York Stock Exchange*). Os resultados apontam que a janela de 15 minutos anterior e posterior às publicações indicam uma reação negativa na movimentação de ambos os índices. Destaca-se, entretanto, que foi aplicada somente a transformação Frequência do Termo – Frequência Inversa dos Documentos (TF-IDF, do inglês *Term Frequency – Inverse Document Frequency*) nos dados textuais, o que pode limitar a extração das relações semânticas entre as palavras presentes no texto. Além disso, a ferramenta de classificação é de uso geral e pode desempenhar imprecisamente em publicações de domínios específicos, como por exemplo, o mercado de ações.

A Tabela I apresenta uma análise comparativa entre os principais pontos de destaque explorados no GOOSE e os recursos utilizados pelos trabalhos mencionados. No geral, a literatura explora a relação entre os sentimentos extraídos das notícias e a movimentação dos índices de mercado. As principais estratégias consistem na rotulagem manual dos dados textuais coletados ou na utilização de ferramentas para a determinação dos sentimentos das notícias. A etapa de rotulagem manual é custosa para o desenvolvimento da pesquisa, e a utilização de ferramentas pode ser limitada de acordo com o contexto da aplicação. Com isso, aproveitar dados devidamente categorizados fornecidos pelas fontes selecionadas é de extrema relevância no desenvolvimento das pesquisas. Por tais motivos, a solução proposta será exposta a seguir.

TABELA I
Comparação de Características com os Trabalhos Relacionados

Trabalho	Crawlers	GloVe Embeddings	Rotulagem Automática
Sert *et al.* [6]	✓		✓
Carosia *et al.* [7]	✓		
Nti *et al.* [8]	✓		✓
Naeem *et al.* [9]			✓
Kinyua *et al.* [10]			✓
GOOSE	✓	✓	✓

III. GOOSE: cateGOrizador de textos curtOS derivados de múltiplas fontEs

Esta seção apresenta o GOOSE [15], uma solução para cateGOrizar textos curtOS derivados de múltiplas fontEs de informações de portais de notícias do mercado financeiro. A etapa inicial envolve a coleta de dados, em seguida do mecanismo de incorporação de palavras e, por fim, o modelo

de classificação para realizar a categorização dos textos em portais de notícias do mercado financeiro.

A. Visão Geral do GOOSE

A Fig. 1 apresenta o fluxograma de execução do GOOSE. A primeira etapa realiza a coleta de dados nos portais de notícias do mercado financeiro. Em seguida, é aplicada a etapa de tratamento de dados em todo o conteúdo textual coletado. Após o tratamento, a terceira etapa executa a criação dos vetores de palavras a partir de todos os dados textuais já tratados e realiza o treinamento da rede Bi-LSTM. Deste total, 80% dos títulos de notícias do portal *Suno Research* são utilizados no treinamento e validação da rede e os 20% restantes para teste. Esta escolha se deve em virtude da consistência de rotulagem das notícias divulgadas nesse portal. Por fim, a etapa de avaliação de resultados é efetivada no conjunto de teste. Os principais componentes de cada etapa são apresentados a seguir.

B. Dataset Modelado

O *dataset* modelado para esta pesquisa foi coletado a partir de três portais de notícias relacionados ao mercado financeiro, sendo eles: (i) *Infomoney* [16]; (ii) *MoneyTimes* [17]; e (iii) *Suno Research* [18]. Os conteúdos das notícias são escritos em português e abordam os seguintes assuntos: negócios, mercado, economia, política e assuntos internacionais.

O *Framework* de código aberto Scrapy [19] foi utilizado para o desenvolvimento dos *crawlers* nos portais de notícias mencionados. A primeira etapa da coleta consiste em armazenar os *links* das notícias em cada página. Em seguida, cada notícia é acessada e são extraídos os seguintes dados: tópico principal, título da notícia, data de publicação, texto completo da matéria, *url* da matéria e tópicos relacionados. Cada portal apresenta uma estrutura HTML padrão, além de realizar diferentes tipos de requisições aos servidores. Observou-se que a estrutura das páginas permaneceu inalterada desde o início das publicações. Com isso, foi possível estabelecer um formato de armazenamento comum a todos os domínios explorados. No entanto, caso exista alteração no formato das páginas, essas mudanças deverão ser implementadas no *crawler* específico de cada portal. A Fig. 2 apresenta uma instância de quais tipos de dados são extraídos e como são armazenados.

A Tabela II apresenta os dados que foram obtidos em cada portal de notícia. De um total de 143.353 notícias coletadas, 34.781 notícias são da *InfoMoney*, 89.234 notícias são da *MoneyTimes* e 19.338 notícias são da *Suno Research*, durante os anos de 2018, 2019 e 2020.

TABELA II
QUANTIDADE DE NOTÍCIAS POR ANO

Fonte	2018	2019	2020	Total
InfoMoney	13.300	10.072	11.409	34.781
MoneyTimes	16.769	31.064	41.401	89.234
Suno Research	1.453	7.712	10.173	19.338

Após a coleta dos textos foi realizada a etapa de tratamento de dados, que consiste na transformação dos textos em letras minúsculas, remoção de pontuações e substituição dos números pelo *token* padrão NUM.

A Fig. 3 apresenta a *WordCloud* gerada a partir da combinação de todos os dados textuais coletados nos portais de notícias e foi gerada após a etapa de preprocessamento dos textos. Com isso, foi possível reduzir o ruído dos dados coletados. Inicialmente, nota-se que as palavras de destaque são *bolsa*, *empresa* e *mercado*, indicando um interesse nesses termos. Em adição, as palavras *investir*, *ações*, *são paulo* e *brasil* reforçam a temática relacionada ao mercado financeiro brasileiro. Os vocábulos *queda*, *alta*, *aumento*, *taxa* e *preço* são relacionados às variações na cotação de ativos, nas tarifas dos impostos e nos índices de mercado.

Os dados textuais do portal *Suno Research* foram selecionados para a execução das etapas de treinamento e avaliação da rede proposta. A escolha foi feita em razão da consistência de classificação das notícias (chave *topic* apresentada na Fig. 2), além da quantidade de exemplos em cada um dos tópicos. As principais categorias selecionadas são apresentados a seguir.

- **Negócios** (8588 publicações): Notícias relacionadas às principais empresas com atuação no Brasil;
- **Mercado** (3643 publicações): Publicações sobre o mercado financeiro, bolsa de valores, ações, câmbio e mercadorias;
- **Internacional** (2378 publicações): Notícias relacionadas à política e à economia internacional;
- **Economia** (2271 publicações): Publicações relacionadas a economia brasileira, tanto aspectos macroeconômicos quanto microeconômicos;

C. Mecanismo de Incorporação de Palavras

Os dados textuais coletados necessitam ter uma representação numérica para serem utilizados como entrada na Bi-LSTM modelada para o GOOSE. Uma das estratégias promissoras é o uso do *GloVe* [20]. *GloVe* recebe um vocabulário com palavras indexadas por inteiros, e associa a cada índice de palavra um vetor n-dimensional. Com o *GloVe* é possível extrair estatísticas globais e relevantes a partir de uma matriz de coocorrência das palavras pertencentes ao corpus fornecido. Com isso, o *GloVe* resolve o problema de dispersão de dados por representar de maneira distribuída as palavras.

Os *GloVe Embeddings* [21] usados no GOOSE foram construídos a partir da concatenação de todos os dados textuais coletados. Foram atribuídos os valores expostos na Tabela III aos parâmetros do algoritmo. A dimensão dos vetores representa o tamanho fixo que cada vetor de palavra será numericamente representado. O número de iterações indica a quantidade de épocas de treinamento do algoritmo. O tamanho da janela representa o contexto associado a quantidade de palavras anteriores e posteriores ao vocábulo em análise, sendo que palavras mais distantes que a janela, terão menor relevância. A contagem mínima indica o mínimo de documentos em que a palavra deve aparecer para ser representada por um vetor.

A visualização dos vetores é realizada a partir da redução de dimensionalidade. A Fig. 4 apresenta a palavra-chave "*banco*", em verde. As palavras em azul e vermelho representam os 20 vetores mais próximos, sendo os pontos em azul os 10

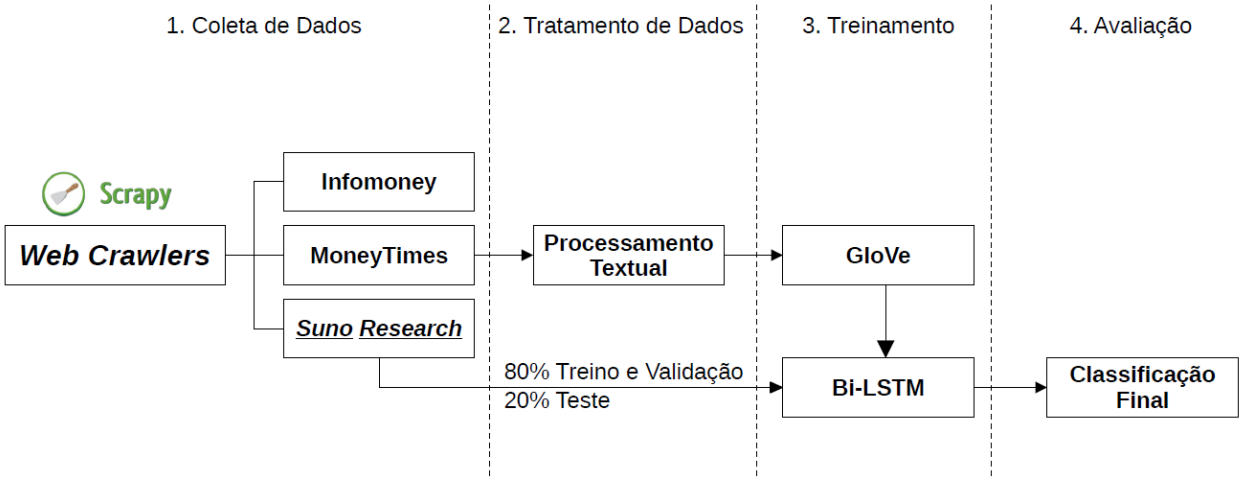

Fig. 1. Fluxo de execução do GOOSE.

```
topic : Mercado
title : Ibovespa fecha em alta de 1,12%, a 119 mil
        pontos, abaixo só do recorde de 23/1
date : 28/12/2020 20:09
full_text : Após a pausa para o Natal, o Ibovespa emendou
            nesta segunda-feira, 28, a terceira sessão
            positiva, que o levou acima dos 119 mil
            pontos nas máximas de hoje, reaproximando o
            índice dos 119.370,48 observados em 18 de
            dezembro, o melhor nível intradia desde 24 de
            janeiro, então no pico histórico de
            119.593,10 pontos. (...)
url  : https://www.sunoresearch.com.br/noticias/ibovespa-
       sobe-119-mil-pontos/
tags [1]
   0  : ibovespa
```

Fig. 2. Formato JSON com dados extraídos de notícia do portal *Suno Research*.

Fig. 3. *WordCloud* dos dados extraídos.

mais relacionados à palavra-chave e os pontos em vermelho os próximos 10 mais associados. Como pode ser observado, as palavras *central*, *bradesco*, *itaú* e *santander* estão entre os vetores mais próximos, pois indicam diretamente o nome das principais instituições bancárias do país. Foram convertidas 9492 palavras em vetores. Para 228 palavras do conjunto de títulos de notícias do portal *Suno Research* não foram

TABELA III
CONJUNTO DE PARÂMETROS SELECIONADOS PARA O MECANISMO DE INCORPORAÇÃO DE PALAVRAS

Parâmetro	Valor
Dimensão dos Vetores	300
Número de Iterações	20
Tamanho da Janela	20
Contagem Mínima	5

encontrados os vetores semânticos correspondentes.

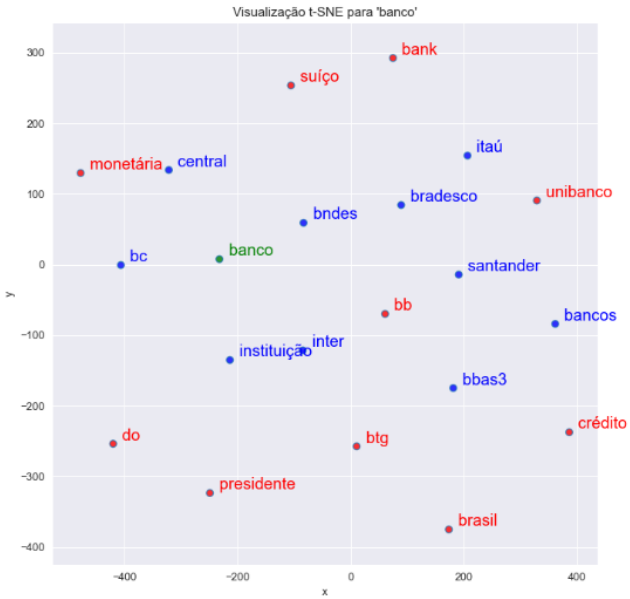

Fig. 4. Vetores de palavras mais relacionados ao vocábulo "*banco*".

D. Arquitetura Modelada

Nesta seção, é apresentada a arquitetura modelada para categorizar os textos curtos derivados de múltiplas fontes

de informações, como ilustrada na Fig. 5. A arquitetura é composta de duas camadas de células LSTM Bidirecionas e camadas *fully connected* intermediária e de saída. Para a camada de saída foi utilizada a função de ativação *softmax* e a função de perda *categorical cross-entropy*, combinação utilizada em problemas de classificação com múltiplos rótulos de saída. Os dados de entrada são vetores semânticos de 300 dimensões, treinados com o algoritmo *GloVe Embeddings* [20], a partir de todos os dados textuais coletados pelos *crawlers*, adotando o mesmo pré-processamento.

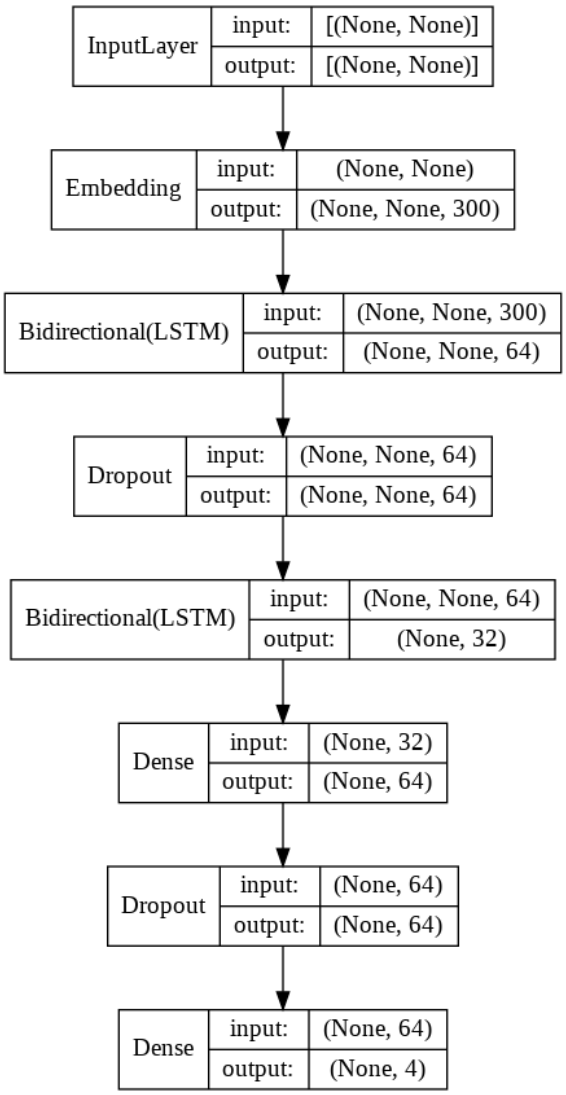

Fig. 5. Bi-LSTM Modelada no GOOSE.

As redes LSTM são modelos particulares de Redes Neurais Recorrentes (RNN, do inglês *Recurrent Neural Network*) e foram desenvolvidas para corrigir o problema encontrado em armazenar informações de curto e longo prazo [22]. As células LSTM (Fig. 6) apresentam canais específicos que têm como funcionalidade filtrar o que será incluído e descartado no estado atual da célula (C_t), descrito pela Eq. 1.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{1}$$

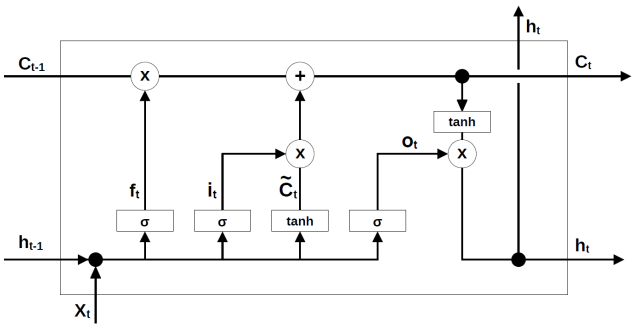

Fig. 6. Célula LSTM.

As informações são codificadas em vetores, os quais sofrem modificações ao passarem por uma sequência de funções. O primeiro portão, indicado pela Eq. 2, faz parte da função de esquecimento da célula (f_t), na qual as saídas próximas a zero indicam que a informação deve ser descartada ao passo que as saídas próximas a 1 indicam que deve ser persistida.

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{2}$$

Em seguida, os próximos portões representados pelas Eqs. (3) e (4), indicam quais valores serão atualizados (i_t) e quais novas informações devem ser armazenadas no estado atual da célula ($\tilde{C}_t$), respectivamente.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = tanh(W_C.[h_{t-1}, x_t] + b_C) \tag{4}$$

A saída da célula, codificada pelo vetor de saída (h_t), será uma versão filtrada do estado atual (C_t), definida pelo produto entre a Eq. 5 e a aplicação da função *tanh* em C_t:

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * tanh(C_t) \tag{6}$$

As redes Bi-LSTM Bidirecionais constituem um modelo posterior às LSTM originais. São modelos de processamento de dados sequências compostos de duas redes LSTM, sendo uma responsável por analisar a entrada para frente e a outra analisa os dados no sentido contrário. Tal organização permite a extração de mais informação a partir dos dados sequenciais.

A aplicação das redes Bi-LSTM para a solução do problema de classificação de relação entre entidades nomeadas em textos foi explorada em [23]. A arquitetura é baseada no presente trabalho. A saída da rede para a representação vetorial da palavra (i) é descrita pela Eq. 7, em que (h_i^f) e (h_i^b) representam, respectivamente, as informações contidas na rede para frente e no sentido contrário.

$$h_i = [h_i^f \oplus h_i^b] \tag{7}$$

IV. Avaliação de Desempenho

Para validar o GOOSE um conjunto de experimentos foi realizado. Para isso, utilizou-se a biblioteca Hyperas [24] para encontrar os melhores hiperparâmetros nos experimentos. O processo de determinação dos hiperparâmetros tem como objetivo otimizar os parâmetros da Bi-LSTM modelada no

GOOSE por meio do *Tree-structured Parzen Estimator* [25]. A Tabela IV apresenta o conjunto de parâmetros utilizados para o processo de hiperparâmetros, sendo que os melhores valores encontrados estão destacados em negrito.

TABELA IV
CONJUNTO DE PARÂMETROS PARA A HIPER-PARAMETRIZAÇÃO

Parâmetro	Valores
#Células na Primeira Camada Bi-LSTM	[8, 16, **32**, 64, 128]
Primeiro *Dropout*	[0 ... **0.29** ... 1]
#Células na Segunda Camada Bi-LSTM	[8, **16**, 32, 64, 128]
#Neurônios na Primeira Camada Densa	[8, 16, 32, **64**, 128]
Função de Ativação	[**tanh**, ReLU]
Otimizador	[Adam, **RMSprop**, SGD]
Segundo *Dropout*	[0 ... **0.73** ... 1]

Para gerar e avaliar a Bi-LSTM, utilizou-se a técnica conhecida como *hold-out* que divide o *dataset* em dois subconjunto, treino e teste. Com isso, o modelo foi treinado com 80% dos títulos do portal *Suno Research*, sendo que, desse total, 20% foi utilizado para teste. Para a análise dos experimentosas seguintes métricas foram utilizadas: (i) precisão; (ii) recall; e (iii) F1-Score. Os experimentos foram realizados no *Google Colab*, ambiente virtual em nuvem com suporte a GPU.

A Fig. 7 apresenta a evolução do erro no conjunto de treinamento e de validação. Com a finalidade de evitar *overfitting* no modelo, foi utilizada a técnica de parada antecipada [26]. O treinamento da rede é interrompido após não identificar diminuição na função de perda em até 15 épocas durante a validação. Como pode ser observado pela curva de aprendizagem, o modelo proposto não possui desempenho drasticamente superior aos dados inéditos. Isso evidência a sua capacidade de generalização na categorização do texto.

A Fig. 8 apresenta o comportamento do valor da acurácia em função da quantidade de épocas. Observa-se que o conjunto de treino minimiza a função de perda e aumenta a acurácia do modelo, e que há uma mudança de comportamento a partir da décima primeira época, assim como na Fig. 7. Isso ratifica que a Bi-LSTM modelada no GOOSE é capaz de resolver o problema de categorização de textos curtos derivados de múltiplas fontes de informações no mercado financeiro.

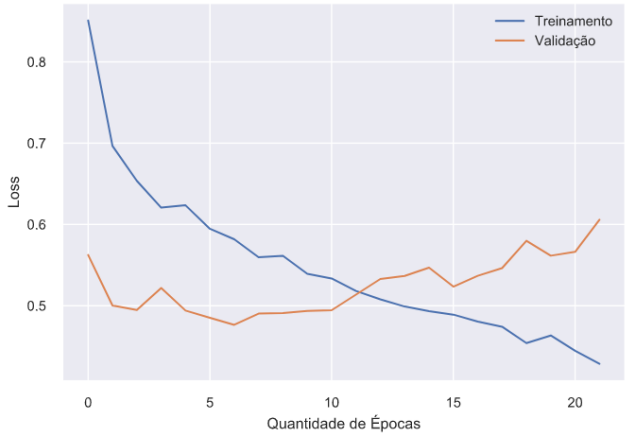

Fig. 7. Gráfico Loss x Quantidade de Épocas.

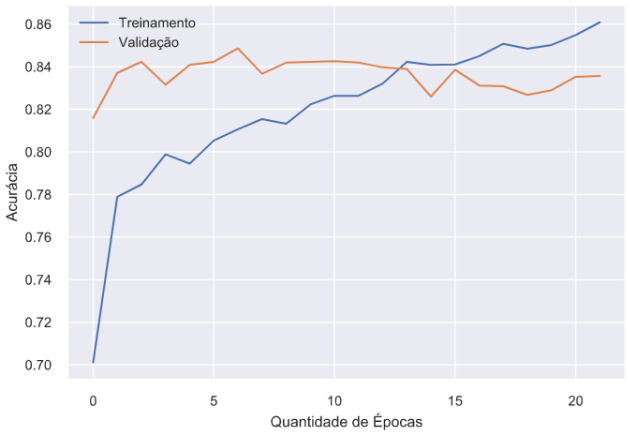

Fig. 8. Gráfico Acurácia x Quantidade de Épocas.

A Fig. 9 apresenta a porcentagem dos resultados obtidos em função das métricas precisão, recall e F1-Score. Independente da categoria extraída dos portais de notícias, observa-se que o GOOSE possui desempenho satisfatório no processo de categorização de textos e na extração da informação. As categorias Negócios e Mercado obtiveram os melhores resultados em relação à métrica F1-Score, atingindo 91% e 82%, respectivamente. É possível observar um aumento no F1-Score conforme as categorias apresentam um maior número de exemplos rotulados. Entre todas as classes, a acurácia média do modelo foi de 84.80%. Isso ocorre em virtude do mecanismo de incorporação de palavras com a camada de atenção da Bi-LSTM que auxilia o GOOSE no processo de categorização de textos curtos.

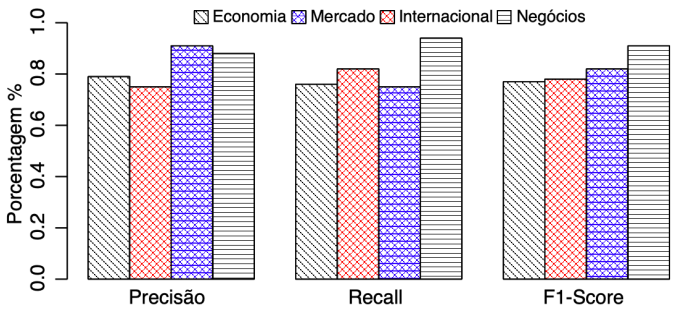

Fig. 9. Impactos de desempenho da Bi-LSTM no GOOSE em relação a precisão, recall e F1-score.

A Fig. 10 apresenta a matriz de confusão da Bi-LSTM modelada em função das quatro categorias. Observa-se que os valores obtidos na diagonal principal da matriz confirmam a aplicabilidade do GOOSE utilizando a rede Bi-LSTM. Tais resultados mostraram que o GOOSE compreende as variações comportamentais dos textos com uma boa eficiência.

V. CONCLUSÃO

Com o objetivo de propor uma solução para a categorização de textos curtos relacionados ao mercado financeiro, este artigo apresentou o GOOSE. Os resultados foram avaliados em um conjunto externo ao treinamento, evidenciando a aplicabilidade do GOOSE para a categorização dos títulos de notícias, sem a

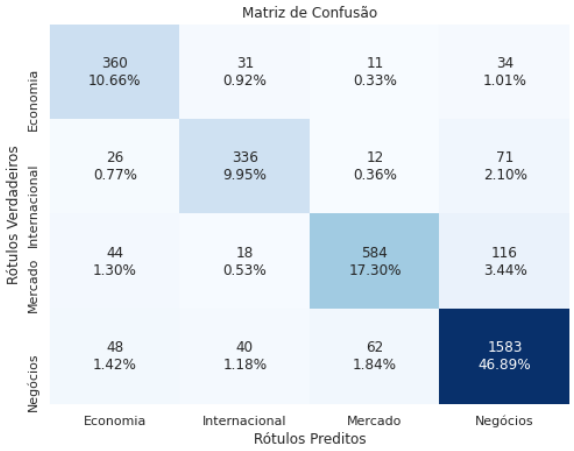

Fig. 10. Matriz de Confusão Final.

necessidade de rotulagem manual dos dados coletados. Ainda, os resultados mostraram que o GOOSE além de apresentar a viabilidade na extração de informações, obteve eficiência no processo de categorização de textos curtos derivados dos portais de notícias do mercado financeiro. Esta pesquisa contribuiu com o desenvolvimento de *crawlers* para três dos principais portais de notícias do mercado financeiro brasileiro: *Infomoney*, *MoneyTimes* e *Suno Research*. Além disso, realizou a criação e análise de vetores semânticos treinados em domínio específico, com os títulos e textos completos de todas as notícias coletadas dos portais. Como trabalhos futuros, planeja-se realizar a coleta de novos dados abertos relacionados ao domínio financeiro para associar com a movimentação do índice Ibovespa, com base em técnicas de Reconhecimento de Entidades Nomeadas e análise de séries temporais.

Referências

[1] http://www.b3.com.br/

[2] S. Sohangir, D. Wang, A. Pomeranets, and T. Khoshgoftaar, "Big Data: Deep Learning for Financial Sentiment Analysis," *Journal of Big Data*, vol. 5, pp. 1–25, 2018.

[3] L. Enamoto, L. Weigang, and G. P. Rocha Filho, "Generic Framework for Multilingual Short Text Categorization Using Convolutional Neural Network," *Multimedia Tools and Applications*, pp. 1–16, 2021.

[4] F. Sebastiani and C. N. D. Ricerche, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, 2002.

[5] C. C. Aggarwal, *Machine Learning for Text*. Springer Publishing Company, Incorporated, 1 ed., (2018).

[6] O. C. Sert, S. D. Şahin, T. Özyer, and R. Alhajj, "Analysis and Prediction in Sparse and High Dimensional Text Data: The Case of Dow Jones Stock Market," *Physica A: Statistical Mechanics and its Applications*, vol. 545, p. 123752, 2020.

[7] 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.

[8] I. k. Nti, A. Adekoya, and B. Weyori, "Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana," *Applied Computer Systems*, vol. 25, pp. 33–42, 06 2020.

[9] M. A. Naeem, S. Farid, B. Faruk, and S. J. H. Shahzad, "Can Happiness Predict Future Volatility in Stock Markets?," *Research in International Business and Finance*, vol. 54, p. 101298, 2020.

[10] J. D. Kinyua, C. Mutigwe, D. J. Cushing, and M. Poggi, "An Analysis of the Impact of President Trump's Tweets on the DJIA and S&P 500 Using Machine Learning and Sentiment Analysis," *Journal of Behavioral and Experimental Finance*, vol. 29, p. 100447, 2021.

[11] K. A. Althelaya, E.-S. M. El-Alfy, and S. Mohammed, "Stock market forecast using multivariate analysis with bidirectional and stacked (lstm, gru)," in *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pp. 1–7, IEEE, 2018.

[12] W. Zhao, G. Zhang, G. Yuan, J. Liu, H. Shan, and S. Zhang, "The Study on the Text Classification for Financial News Based on Partial Information," *IEEE Access*, vol. 8, pp. 100426–100437, 2020.

[13] http://hedonometer.org/timeseries/en_all

[14] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," The AAAI Press, 2014.

[15] https://github.com/mso13/BrazilianFinancialNews

[16] https://www.infomoney.com.br/

[17] https://www.moneytimes.com.br/

[18] https://www.sunoresearch.com.br/noticias/

[19] https://scrapy.org/

[20] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[21] https://github.com/stanfordnlp/GloVe/

[22] S. Hochreiter and Schmidhuber, J., "Long Short-Term Memory," *Neural Comput.*, vol. 9, pp. 1735–1780, November 1997.

[23] S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional Long Short-Term Memory Networks for Relation Classification," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, (Shanghai, China), pp. 73–78, Oct. 2015.

[24] https://github.com/maxpumperla/hyperas

[25] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for Hyper-Parameter Optimization," in *Advances in neural information processing systems*, pp. 2546–2554, 2011.

[26] L. Prechelt, "Early Stopping-But When?," in *Neural Networks: Tricks of the Trade* (G. B. Orr and K.-R. Müller, eds.), vol. 1524 of *Lecture Notes in Computer Science*, pp. 55–69, Springer, 1996.

Matheus Schmitz (matheus.schmitz@aluno.unb.br) é graduado em Engenharia de Computação (2019) pela Universidade de Brasília (UnB), Brasil. Atualmente é mestrando no Programa de Pós-Graduação em Informática (PPGI) vinculado ao Departamento de Ciência da Computação (CiC) da UnB. Tem interesse nas linhas de pesquisa relacionadas ao Aprendizado de Máquina e ao Processamento de Linguagem Natural.

Roger Immich é Professor do Instituto Metrópole Digital (IMD) da Universidade Federal do Rio Grande do Norte (UFRN). Ele recebeu seu Ph.D. em Engenharia Informática pela Universidade de Coimbra, Portugal (2017). Foi pesquisador visitante na Universidade da Califórnia em Los Angeles, Estados Unidos (UCLA) em 2017, e realizou pós-doutorado no Instituto de Computação da Universidade de Campinas (UNICAMP) em 2019. Seus interesses de pesquisa são em Smart Cities, IoT, Quality of Experience, bem como Cloud and Fog computing.

Gustavo Pessin é Pesquisador do Instituto Tecnológico Vale, Mineração. Pessin obteve seu o título de Doutor em Ciência da Computação pela Universidade de São Paulo, como membro do *Mobile Robotics Lab*. Durante seu doutorado Pessin desenvolveu pesquisas no *Robotics Lab*, na *Heriot-Watt University*, Edimburgo, Reino Unido, e no *Communication and Distributed Systems Group*, na Universität Bern, Suíça. Em 2015, Pessin ocupou o cargo de *Visiting Scholar no Media Lab do Massachusetts Institute of Technology*. Suas pesquisas são relacionadas com robótica móvel autônoma, aplicações com aprendizado de máquina, data analytics e IoT industrial.

Geraldo P. Rocha Filho (geraldof@unb.br) é Professor adjunto do Departamento de Ciência da Computação da Universidade de Brasília. Foi Pesquisador no Instituto de Computação da UNICAMP por meio do Pós-Doutorado em 2018. Obteve o título de Doutor e Mestre em Ciência da Computação e Matemática Computacional pelo ICMC-USP em 2018 e 2014, respectivamente. Seus interesses de pesquisa são redes de sensores sem fio, redes veiculares, redes inteligentes, cidades inteligentes e aprendizado de máquina.