Using a labeling function for automatic classification of agribusiness news: A weak supervisory approach

Rodrigo Neves Trindade[1], **Luiz H. D. Martins**[1],
Geraldo Nunes Correa[1], **Ivan José dos Reis Filho**[1]

rodrigo.1093795@discente.uemg.br, luiz.1093701@discente.uemg.br

geraldo.correa@uemg.br, ivan.filho@uemg.br

[1]Departamento de Ciências Exatas – Universidade do Estado de Minas Gerais (UEMG) – Frutal, MG – Brazil

Abstract. *The large volume of news generated on the internet has increased the use of machine learning applications. Predictive models need labeled samples in large quantity and quality to ensure reasonable accuracy in classification tasks. However, labeling news is a manual task and time-consuming for the domain expert. In this work, a function is proposed to label agribusiness news. Soybean price series oscillations in the domestic and international markets and the dollar quotation are input for the labeling function. Different learning paradigms and textual representations are used in the assessment stage. The neural language models showed better performance, indicating that the proposal can be an alternative for real-time applications.*

Resumo. *O grande volume de notícias geradas na internet têm aumentado o uso de aplicações com aprendizado de máquina. Modelos preditivos precisam de amostras rotuladas em grande quantidade e qualidade para garantir boa acurácia em tarefas de classificação. No entanto, a tarefas de rotular notícias é manual e demanda tempo do especialista de domínio. Neste trabalho, uma função é proposta para rotular notícias do agronegócio. Oscilações das séries de preços da soja no mercado nacional, internacional e cotação do dólar são a entrada para a função de rotulagem. Diferentes paradigmas de aprendizado e representações textuais são usadas na etapa de avaliação. Os modelos de linguagem neural demonstraram melhor desempenho e os resultados indicam que a proposta pode ser uma alternativa para aplicações de tempo real.*

1. Introdução

Devido ao grande volume de dados gerados diariamente na internet, aplicações computacionais tem crescido exponencialmente [Ratner et al. 2020]. O avanço tecnológico na área da computação têm possibilitado a transformação de dados em informações úteis, visando auxiliar os investidores, os produtores e as empresas na sua tomada de decisão. Novas aplicações que usam recursos de aprendizado de máquina têm facilitado o apoio necessário para reduzir custos computacionais a fim de armazenar, analisar e gerenciar um grande volume de dados [Chatfield and Xing 2019].

Aplicações que usam métodos de aprendizado de máquina utilizam um grande volume de dados para aprender com base em exemplos. Entretanto, a qualidade dos resultados dos modelos preditivos dependem que as amostras sejam rotuladas e disponibilizadas

em grande quantidade e qualidade [Boecking et al. 2020]. Um dos problemas enfrentados por especialistas do domínio é a tarefa de rotular uma quantidade suficiente de amostras, em tempo de serem utilizados em aplicações de tempo real [Zhou 2018]. Por exemplo, diariamente um grande volume de notícias são disponibilizadas na internet referente ao mercado financeiro. Contudo, devido a natureza dinâmica e complexa do domínio, amostras precisam ser rotuladas frequentemente para representar fatores recentes e não lineares para complementar o conjunto de amostras.

Uma alternativa para rotular um grande volume de notícias é utilizar técnicas de supervisão fraca [Zhou 2018]. Técnicas de supervisão fraca podem ser uma possibilidade para rotulagem automática de notícias em tempo real e treinar modelos sem a necessidade da ajuda humana [Boecking et al. 2020]. Estudos recentes têm usado técnicas em diferentes tipos de aplicações, como a rotulação automática de notícias para detecção de *fake news* [Wang et al. 2020], rotulação de imagens com o auxílio de mídias sociais [Dai et al. 2021], reconhecimento de entidades nomeadas [Lison et al. 2020] e classificação de textos usando fontes externas [Ratner et al. 2020].

A tarefa de rotular notícias pode ser feita diariamente, utilizando fontes de dados externos, mas que representam o mesmo domínio. Por exemplo, oscilações abruptas da série de preços do mercado financeiro pode ser ocasionadas por eventos contidos em informações de textos [Munezero et al. 2014]. Nesse sentido, este trabalho propõe uma função para rotular notícias de modo automático, com base nas oscilações de três séries temporais. A variação do preço da soja praticado internacionalmente no *Chicago Board of Trade* (CBOT), a série de preço praticado no porto brasileiro e a cotação do dólar, são usados como entrada para a função de rotulagem. Seis representações vetoriais de textos e cinco modelos preditivos são utilizados a fim de avaliar o desempenho da classificação. Os resultados são apresentados e discutidos na seção específica de avaliação.

Com o intuito de apresentar a abordagem proposta neste trabalho, as seções são divididas da seguinte forma: Seção 2 traz a discussão de trabalhos relacionados; Seção 3 os métodos utilizados; Seção 4 a avaliação dos métodos utilizados; Seção 5 os resultados obtidos no trabalho; Seção 6 denota a discussão do trabalho; e por fim a Seção 7 a conclusão obtida do trabalho.

2. Trabalhos relacionados

Modelos preditivos que aprendem utilizando abordagens de supervisão fraca são caracterizadas como: supervisão incompleta, inexata e imprecisa [Zhou 2018]. A supervisão incompleta refere-se ao subconjunto de dados de treinamento rotulados, enquanto o restante dos dados permanece sem a rotulagem. A inexata, em que os dados de treinamento fornecem uma rotulagem de granulação grossa, e por último, a imprecisa, em que, geralmente os dados não são verdadeiros. A supervisão fraca utilizam tentativas de rotulagem automática, gerando assim rótulos mais fracos com base no domínio [Ratner et al. 2020].

Na supervisão incompleta existe várias tarefas que podem ser aplicadas, um exemplo é o aprendizado semi-supervisionado, no qual apenas um subconjunto de dados são rotulados. Modelos discriminativos e híbrido generativo são métodos bem recebidos em tarefas de aprendizado de máquina [Zhu 2005][Lasserre et al. 2006]. No entanto, algumas versões de *autoencoders* também foram utilizados na tarefa semi-supervisionada que utiliza suas camadas com propriedades generativas e a discriminação das diferentes amos-

tras conforme as suas evidências.

Em relação a supervisão inexata, alguns trabalhos utilizam rótulos previamente fornecidos, mas não apresentam uma precisão desejada [Anklin et al. 2021][Dai et al. 2021]. Na abordagem de supervisão imprecisa estudos têm proposto um método heurístico aprendendo com o *feedback* do usuário, demonstrando que é possível em apenas algumas interações o modelo treinado alcançar um desempenho relativamente competitivo sem ter acesso aos rótulos de treinamento [Boecking et al. 2020].

Para o tratamento na identificação de notícias falsas foram utilizados dois conjuntos de dados e seis modelos de representação textual no estudo [Helmstetter and Paulheim 2021]. Tais notícias foram extraídas da rede social *Twitter*, sendo rotuladas por meio de suas fontes. Para as notícias falsas os autores levou em consideração perfis conhecidos por espalhar *fake news* e rotulou como precisos *tweets* emitidos por perfis confiáveis. Sendo exploradas duas alternativas: um modelo *Bag-of-Words* (BoW), utilizando o modelo Frequência de Termo - Frequência de Documento Inverso (TF-IDF) e o modelo neural Doc2Vec. Contudo, os autores optaram pela criação de um conjunto de dados em maior escala, possuindo os rótulos imprecisos e assim gerando resultados competitivos.

Para a classificação de textos curtos com dados não rotulados e uma diferença desequilibrada do espaços dos dados, foi utilizado no trabalho [Chen et al. 2022]. O método proposto pelos autores geram rótulos probabilísticos por meio da independência condicional, sendo adotado os modelos de pré-treinamento: Bert-Base e Multilingual Chinês, RoBERTa e Chinês Grande, Chinês ERNIE e ERNIE. Todavia, o estudo mostra que dados não rotulados desequilibrados para problemas de classificação podem ser resolvidos com múltiplas formas de supervisão fraca.

3. Métodos

Este trabalho propõe uma abordagem para rotular notícias de uma *commoditie* agrícola. Séries de preços da soja, praticados no CBOT, no porto brasileiro (CEPEA) e variações do dólar, são utilizados como entrada para a função de rotulagem. A Figura 1 apresenta as etapas realizadas neste trabalho.

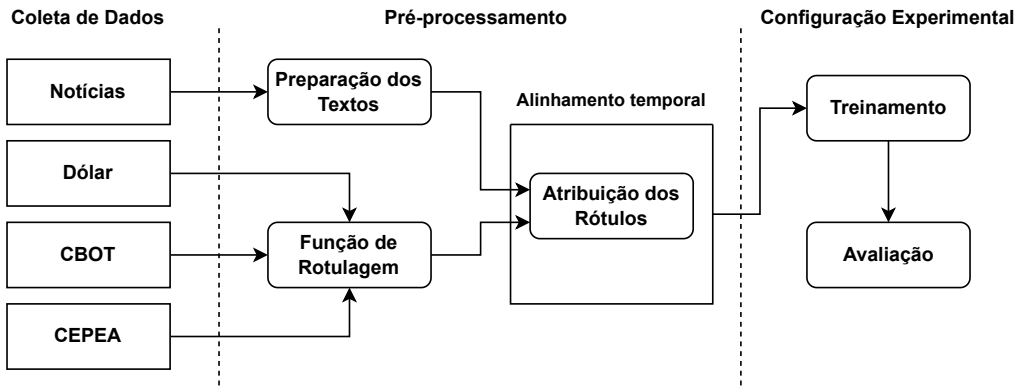

Figura 1. Representação conceitual da abordagem proposta

Na primeira etapa de **Coleta de Dados**, representa a obtenção de dados textuais e séries de preços da *commodities* agrícola. A etapa de **Pré-processamento** é feito a preparação dos textos e aplicada a função de rotulagem. Na última etapa de **Configuração experimental** é realizada as etapas de treinamento de modelos preditivos e avaliação dos resultados. As configurações realizadas em cada etapa são apresentadas a seguir.

3.1. Função de rotulagem

Na etapa de pré-processamento dos dados, inicialmente verificou-se a diferença dos valores intra dia nas séries de preços. Conceitualmente, uma série de preço é definida S de tamanho m com uma sequência ordenada das observações S=(S_1, S_2, S_3, ..., S_m), em que s_t representa uma sequência s no tempo t. Para os textos também define uma sequência ordenada $D = (D_1, D_2, D_3, ..., D_m)$, em que d_t é um documento d no tempo t. Em seguida, realizou um alinhamento temporal entre as notícias e as séries de preço. Os rótulos (0 e 1) são atribuídos aos documentos em relação a diferença intra-dia s_p (de um dia para o outro) com o limiar p. A Equação 1 apresenta a função de rotulagem.

$$d_t = \begin{cases} 0 & if \quad s_p \quad < \quad p \\ 1 & if \quad s_p \quad \geq \quad p \end{cases} \tag{1}$$

Se a oscilação intra-dia é maior ou igual ao limiar estabelecido p, o rótulo 1 é atribuído a notícia, caso contrário é rotulado como 0. Quanto maior o valor de p, indica uma oscilação abrupta intra-dia na série de preço. Nesse sentido, consideramos a premissa que o conteúdo da notícia em dias que tem oscilações abruptas, potencialmente contém informações relevantes e que motivaram na oscilação do preço. Na etapa de **atribuição de rótulos**, o método verifica a maior ocorrência dos *labels* extraídos da função de rotulagem do Dólar, CBOT e CEPEA. Por exemplo, se for atribuído rótulo 1 para oscilação do dólar, rótulo 1 para oscilação do CBOT e rótulo 0 para o CEPEA, a notícia recebe rótulo 1 devido a maior ocorrência da saída da função de rotulagem. A Tabela 1 apresenta exemplos do resultado da função de rotulagem.

Manchete	**Dólar**	**CEPEA**	**CBOT**	**rótulo**
Soja sobe no interior do Brasil, apesar do fechamento negativo na CBOT nesta 5ª feira	0	1	0	0
Soja tem mais uma sessão de estabilidade em Chicago na manhã desta 5ª feira	0	1	1	1
Greve dos caminhoneiros para o transporte até os portos e pressiona preços no Brasil	0	0	0	0
Safra de soja eleva PIB da agropecuária no 1º tri e deve garantir alta no ano	1	1	1	1

Tabela 1. Método para atribuir rótulo

A Tabela 1 apresenta manchetes de notícias que foram rotuladas como 0 ou 1, sendo atribuído o rótulo $(0, 1)$ de acordo com a maior ocorrência dos rótulos do dólar, CEPEA e CBOT. Entre as 1.800 manchetes extraídas (Tabela 3), 925 foram rotuladas como positivas (1) e 876 manchetes como negativas ou neutras (0). Nota-se que fatores como questões de transportes, notícias relacionadas com a Bolsa de Chicago e economia tendem a influenciar no mercado interno (Tabela 1). Dessa forma, reforça a hipótese que a oscilação de preços dos principais meios de negociação nacional e internacional pode ser uma alternativa para rotular notícias de modo automático.

3.2. Preparação dos Textos

Algoritmos que utilizam aprendizado de máquina para classificação de textos, exigem uma estrutura vetorial adequada para as etapas de treinamento e teste. Com o propósito de construir uma representação estruturada é necessário aplicar o pré-processamento dos dados. As principais formas de representação textual são baseadas na frequência do termo *Bag-of-Words* (BoW) e os modelos de linguagem neural.

Para o processamento dos textos foram realizadas tarefas de tratamento, limpeza e redução no volume de dados textual para gerar vetores que representam textos. Para atividades serem feitas de forma correta foram realizadas algumas técnicas, tais como: *i)* remoção das *stopwords*, visando eliminar alguns artigos, preposições, pronomes e conjunções que não trazem informação relevante ao contexto do documento; *ii)* a normalização, cujo objetivo é eliminar variações que a palavras vem assumir. A estrutura da *Bag-of-Words* normalmente pode ser representada por um modelo de espaço vetorial, onde as palavras são indexadas e ponderadas por medidas de ocorrência entre termos e documentos [Aggarwal and Reddy 2014]. A Tabela 2 apresenta a estrutura vetorial da BoW.

	t_1	t_2	t_3	$\cdots$	t_n
d_1	Wd_1,t_1	Wd_1,t_2	Wd_1,t_3	$\cdot$	Wd_1,t_n
d_2	Wd_2,t_1	Wd_2,t_2	Wd_2,t_3	$\cdot$	Wd_2,t_n
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
d_m	Wd_m,t_1	Wd_m,t_2	Wd_m,t_3	$\cdot$	Wd_m,t_n

Tabela 2. Representação Matriz de *m* documentos e *n* atributos

A Tabela 2 ilustra o modelo de espaço vetorial para uma coleção com *m* documentos e *n* atributos, onde $D = [d_1, d_2, d_3, ..., d_m]$ representa o conjunto de documentos e $T = [t_1, t_2, t_3, ..., t_n]$ o conjunto de atributos. A representação é chamada de saco de palavras, pois as dimensões vetoriais representa termos simples [Aggarwal 2018]. Os valores dos pesos são calculados na frequência dos termos no documento, podendo ser utilizadas as medidas mais comuns, como: i) binária, onde é representada de forma binária como 0 e 1, 0 como ausência e 1 como presença do termo no documento; ii) Frequência do Termo (*Term Frequency* (TF)) que define a frequência do termo no documento; iii) Frequência de Termo - Frequência de Documento Inverso (*Term Frequency-Inverse Document Frequency* (TF-IDF)) analisa o TF com a sua frequência inversa do termo no documento. Os modelos utilizados para avaliação são Binário, TF e TF-IDF.

Modelos de linguagens neurais foram projetados para representações textuais que consideram recursos e estrutura semânticas, tal como o *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al. 2018]. Atualmente existem modelos pré-treinados alternativos, cujo objetivo é melhorar o modelo e diminuir custo computacional, por exemplo: RoBERTa, modelo que utiliza sequências maiores que o modelo original, remove a seguinte meta de previsão da sentença e altera dinamicamente o padrão do mascaramento [Liu et al. 2019]; outro modelo é o DistilBERT, cujo a estratégia é economizar memória ficando mais rápido que o BERT original, pois na fase de pré-treinamento utiliza-se um menor número de parâmetros, técnica de destilação e a técnica de perda tripla [Sanh et al. 2019]; o modelo BERTimbau é um modelo pré-treinado com palavras

da língua portuguesa treinado por BrWac (Brazilian Web as Corpus)[Souza et al. 2020]. Os modelos pré-treinados com base na arquitetura BERT geram vetores 768 ou 1024 posições. Contudo, os modelos usados no trabalho utilizaram os vetores de 768 posições, tais como BERT, DistilBERT e BERTimbau.

4. Avaliação

Este trabalho apresenta uma avaliação de supervisão fraca usando um conjunto de dados de uma commoditie agrícola. Além disso, uma comparação entres os modelos preditivos foi realizada para classificação de textos, utilizando diferentes modelos de representação textual. Diferentes paradigmas de aprendizado de máquina foram utilizados para avaliar o desempenho de classificação do método proposto. O método KNN (*K-Nearest Neighbors*) é o paradigma baseado em instância. O método MLP (*Multi-Layer Perceptron*) baseado no paradgima conexionista. No método de paradigma probabilístico utilizou-se o modelo GNB (*Gaussian Naive Bayes*). Do paradigma estatístico valeu-se do modelo SVM (*Support Vector Machine*). E, por fim, o modelo DTC (*Decision Tree Classifier*) representa o paradigma simbólico.

4.1. Coleta dos Dados

Dados textuais referentes a *commodities* de soja foram extraídos do site *Notícias Agrícolas* [1]. Fundado em 1997, atualmente é um dos mais importantes meios de comunicação no agronegócio brasileiro. Diariamente são publicadas de duas à cinco notícias de diferentes assuntos relacionados à *commodities*. Neste trabalho, foi considerada uma notícia por dia devido grande variedade disponível no *site*. Dessa forma, definiu-se uma lista de palavras relacionadas aos fatores importantes na formulação dos preços da *commodities*, tais como: política, transporte, exportação, doença e pragas, nutrição animal e clima. Dessa forma, foi escolhido a notícia que continha maior número de ocorrências das palavras previamente estabelecidas. A Tabela 3 descreve o período do conjunto de dados, o número de dias e os atributos das séries temporais.

Período	02/01/2015 à 25/05/2022
Números de dias	1800
Atributos ST	Fechamento do valor diário (dólar, cbot, cepea)

Tabela 3. Visão geral dos dados textuais

Os dados das séries temporais do dólar e valores negociado na bolsa de valores de Chicago CBOT foram extraídas da plataforma *Macrotrends*[2], principal plataforma de pesquisa para investidores de longo prazo, e o valor da *commodity* negociado na bolsa de valores B3, extraídos do Centro de Estudos Avançados em Economia Aplicada (*CEPEA*[3]) na Universidade de São Paulo (USP).

4.2. Configuração experimental

Nesta etapa foram utilizados cinco algoritmos de classificação tradicionais: KNN, MLP, GNB, SVM e DTC. Os parâmetros utilizados foram os valores padrão da biblioteca *scikit-learn* [Pedregosa et al. 2011]. A estratégia de avaliação para o **treinamento** foi a divisão

[1] https://www.noticiasagricolas.com.br/noticias/soja/
[2] https://www.macrotrends.net/
[3] https://www.cepea.esalq.usp.br/br/indicador/soja.aspx

hold-out, dividindo o conjunto de dados de 80% para treinamento (1.440 notícias) e 20% para teste (360 notícias). Esse método geralmente é usado quando se tem um conjunto de dados maior ou está no início da construção de um modelo, exigindo assim um poder computacional menor. Além disso, a técnica permite treinar notícias temporalmente do passado para classificar notícias do futuro. Na etapa de **avaliação** dos resultados utilizou a média harmônica *F1-Score*, métrica que considera a combinação entre precisão (*precision*) e revoação (*recall*).

5. Resultados

A avaliação experimental investiga abordagem da supervisão fraca por meio da função de rotulagem proposta (Equação 1), objetivando analisar o impacto de cada representação textual e o modelo preditivo. A Tabela 4 apresenta os resultados obtidos da classificação dos algoritmos MLP, SVM, KNN, GNB e DTC com o conjunto de dados da soja, sendo atribuído o rótulo (1) ao texto, obtendo uma variação da diferença dos preços.

Modelo	**Binário**	**TF**	**TF-IDF**	**BERT**	**DistilBERT**	**BERTimbau**
MLP	0.587	**0.596**	0.583	0.515	0.544	0.532
SVM	**0.430**	0.429	0.428	0.419	0.418	0.419
KNN	**0.540**	0.519	0.505	0.523	0.480	0.485
GNB	0.480	0.480	0.482	**(0.630)**	0.605	0.572
DTC	0.529	0.536	0.520	**0.588**	0.475	0.526

Tabela 4. Resultados da avaliação. Comparação (macro F1) de BoW e linguagem neural pré-treinados

Conforme apresentado na Tabela 4, os valores em negrito representam os melhores resultados de cada modelo preditivo e os valores sublinhados refere ao melhor resultado de cada representação textual. O modelo MLP obteve o melhor resultado com a representação TF no valor de 0.596, o modelo SVM obteve o valor de 0.430 com a representação *Binary*, o modelo KNN também obteve um maior valor de 0.540 com a representação *Binary*, e, por fim, o modelo DTC obteve o valor de 0.588 com a representação BERT. O melhor resultado entre todos os valores é o GNB com o valor de 0.630 com a representação BERT. A Tabela 5 apresenta os resultados das métricas de avaliação *precision*, *recall*, *f1-score* e *support* do melhor resultado obtido na Tabela 5.

	Precision	**Recall**	**F1-score**	**Suporte**
0	0.800	0.757	0.778	260
1	0.452	0.514	0.481	101
accuracy	0.689			
macro avg	0.626	0.636	**0.630**	361
weighted avg	0.703	0.689	0.695	361

Tabela 5. Métrica de avaliação do melhor resultado da classificação

Conforme apresentado na Tabela 5 a precisão o resultado de 0.689. No entanto, nota-se que rótulos fracos estão desequilibrados. Contudo, realizando uma análise dos resultados para esse tipo de avaliação, a média harmônica *F1-Score* se torna a mais apropriada. Os valores de suporte representa os 20% da etapa de teste.

6. Discussão

De acordo com os resultados apresentados na Tabela 4, a representação TF (0,596) obteve o melhor valor de *F1-Score* com o modelo MLP entre as representações baseadas na BoW. O modelo MLP se destacou entre as representações BoW, obtendo os melhores resultados. Os resultados obtidos por meio dos modelos de linguagens neurais, destaca-se o modelo preditivo GNB com melhor resultado para a representação BERT (0.630). A Figura 2 ilustra um gráfico de todos os resultados obtidos na etapa de classificação (Tabela 4).

Figura 2. Gráfico dos resultados obtidos com a classificação

Em relação ao modelo MLP, observa-se que os resultados tem valores semelhantes, sendo a representação TF com o melhor resultado. O modelo SVM obteve o pior desempenho entre todas as representações textuais com valores abaixo de 45% de F1 Score. O modelo KNN obteve valores intermediários em relação aos piores e melhores resultados. O modelo GNB obteve melhor desempenho para as representações DistilBERT e BERT. Por fim, a representação DistilBERT obteve o pior desempenho entre todas representações para o modelo DTC. A Figura 3 apresenta uma avaliação estatística de *Nemenyis* entre os melhores resultado obtidos na Tabela 4 .

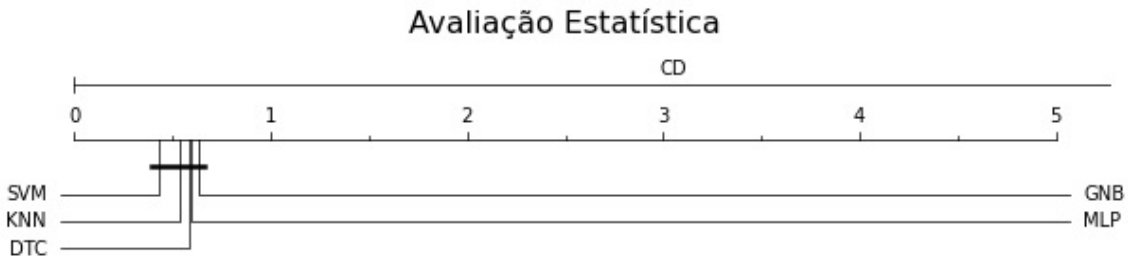

Figura 3. Diagrama da diferença crítica de Friedman com o pós-teste de Nemenyis

Conforme ilustra a Figura 3, uma comparação estatística dos modelos preditivos, com base no teste não paramétrico de *Friedman* juntamente com o pós-teste de *Nemenyis* por meio de um diagrama de teste crítico da diferença [García et al. 2010]. Contudo, nota-se que os modelos preditivos utilizados não houve uma diferença significativa no teste.

7. Conclusão

Este trabalho apresentou uma abordagem de supervisão fraca para rotulação de notícias usando oscilações de séries temporais. Manchetes de notícias relacionadas a *commodities* de soja foram rotuladas com base nas oscilações da séries de preços praticados na CBOT, Dólar e CEPEA. Na avaliação experimental considerou-se diferentes tipos de paradigmas de aprendizado e seis representações textuais para avaliar o desempenho da tarefa de classificação.

Analisando os resultados pode-se observar que a rotulagem de um grande volume de notícias por meio da função de rotulagem proposta pode ser uma alternativa em aplicações de tempo real. Considerando os melhores resultados na configuração experimental, a representação de linguagem neural BERT e o modelo preditivo GNB alcançaram F1-score de 0.630 e acurácia de 0.689. Além disso, os resultados indicam que os modelos de linguagens neurais são melhores alternativas para cenários de avaliação de supervisão fraca.

Uma limitação da avaliação são as classes desbalanceadas e estratégia de avaliação utilizada. Como trabalhos futuros podem ser considerada a estratégia de *time-serie split* e mais dados externos para auxiliar na rotulação de notícias.

Agradecimentos: Os autores agradecem ao Centro Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (Processo PCRH BPG-00054-210).

Referências

Aggarwal, C. C. (2018). *Machine learning for text*, volume 848. Springer.

Aggarwal, C. C. and Reddy, C. K. (2014). Data clustering. *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra*.

Anklin, V., Pati, P., Jaume, G., Bozorgtabar, B., Foncubierta-Rodriguez, A., Thiran, J.-P., Sibony, M., Gabrani, M., and Goksel, O. (2021). Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 636–646. Springer.

Boecking, B., Neiswanger, W., Xing, E., and Dubrawski, A. (2020). Interactive weak supervision: Learning useful heuristics for data labeling. *arXiv preprint arXiv:2012.06046*.

Chatfield, C. and Xing, H. (2019). *The Analysis of Time Series: an introduction with R*. CRC press.

Chen, L.-M., Xiu, B.-X., and Ding, Z.-Y. (2022). Multiple weak supervision for short text classification. *Applied Intelligence*, 52(8):9101–9116.

Dai, E., Shu, K., Sun, Y., and Wang, S. (2021). Labeled data generation with inexact supervision. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 218–226.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

García, S., Fernández, A., Luengo, J., and Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information sciences*, 180(10):2044–2064.

Helmstetter, S. and Paulheim, H. (2021). Collecting a large scale dataset for classifying fake news tweets using weak supervision. *Future Internet*, 13(5):114.

Lasserre, J. A., Bishop, C. M., and Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 87–94. IEEE.

Lison, P., Hubin, A., Barnes, J., and Touileb, S. (2020). Named entity recognition without labelled data: A weak supervision approach. *arXiv preprint arXiv:2004.14723*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Munezero, M., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2020). Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2):709–730.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Wang, Y., Yang, W., Ma, F., Xu, J., Zhong, B., Deng, Q., and Gao, J. (2020). Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 516–523.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53.

Zhu, X. J. (2005). Semi-supervised learning literature survey.