Classificação Multi-classe para Análise de Qualidade de Feedback

Hyan H. N. Batista[1], **Anderson Pinheiro Cavalcanti**[2,4], **Péricles Miranda**[1], **André Nascimento**[1],**Rafael Ferreira Mello**[1,3]

[1] Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)

[2]Centro de Informática – Universidade Federal de Pernambuco (UFPE)

[3]Centro de Estudos e Sistemas Avançados do Recife (CESAR School)

[4]SiDi Recife

{rafael.mello,pericles.miranda,andre.camara}@ufrpe.br

Abstract. *Feedback is a very important factor in the teaching-learning process and crucial in Distance Education, because, as teachers and students are separated in space and/or time, it is through feedback that the student will understand how their performance is in the classroom. discipline and what are the next steps of learning. There are feedback models in the literature that help the teacher to structure and provide quality feedback to the student. In this work, we use Hattie and Timperley's highly regarded feedback model, which divides feedback into categories (task, task processing, regulation, and personal). It is possible to find in the literature works that analyze feedback automatically based on this model. However, these works use traditional machine learning algorithms and train binary classifiers for each level of feedback. Thus, this work aims to use deep learning algorithms for multi-class feedback classification based on the Hattie and Timperley model.*

Resumo. *O feedback é um fator muito importante no processo de ensino-aprendizagem e crucial na Educação a Distância, pois, como professores e alunos estão separados no espaço e/ou tempo, é através do feedback que o aluno vai entender como está o seu desempenho na disciplina e quais são os próximos passos do aprendizado. Existem na literatura modelos de feedback que ajudam o professor a estruturar e fornecer um feedback de qualidade ao aluno. Nesse trabalho utilizamos o conceituado modelo de feedback de Hattie e Timperley que divide o feedback em categorias (tarefa, processamento da tarefa, regulação e pessoal). É possível encontrar na literatura trabalhos que analisam feedback automaticamente com base nesse modelo. Contudo, esses trabalhos utilizam algoritmos tradicionais de aprendizagem de máquina e treinam classificadores binários para cada nível de feedback. Dessa forma, este trabalho tem como objetivo utilizar algoritmos de deep learning para classificação multi-classe de feedback com base no modelo de Hattie e Timperley.*

1. Introdução

O feedback foi identificado como um dos dez principais aspectos da aprendizagem para melhorar o desempenho do aluno [Hattie and Gan 2011]. De acordo com [Sadler 1989],

DOI: 10.5753/sbie.2022.225396

o feedback precisa fornecer informações relevantes relacionadas a uma tarefa ou processo de aprendizagem e evitar discrepâncias entre o conhecimento adquirido pelo aluno e o que a disciplina deveria ensinar. Além disto, [Laurillard 1993] afirma que o ensino sem feedback é completamente improdutivo para o aluno. Os princípios de boas práticas de feedback reconhecem que o feedback não é um produto, mas um processo complexo que deve conscientizar os alunos sobre como o comportamento, as emoções e a cognição do estudo real influenciam seus resultados [Boud and Falchikov 2007, Henderson et al. 2019].

Segundo [Ypsilandis 2002], o feedback é um fator crucial para o sucesso ou o fracasso de um curso à distância, pois esses cursos geralmente têm uma alta taxa de desistência, e também pelo fato de alunos, professores e tutores estarem separados fisicamente. Assim, interações e feedback informativos e oportunos se tornam ainda mais críticos para a construção do conhecimento e o sucesso acadêmico [Joulani et al. 2013].

Em ambientes virtuais de aprendizagem, os professores geralmente propõem diversas atividades onde os alunos enviam respostas e recebem feedback sobre seu progresso [Coates et al. 2005]. No entanto, é um desafio para os professores fornecer feedback informativo e de alta qualidade devido ao número crescente de alunos matriculados em cursos online. Uma característica importante que vem sendo analisada nas últimas décadas é a qualidade do feedback [Hattie and Timperley 2007, Nicol and Macfarlane-Dick 2006]. Por exemplo, [Cavalcanti et al. 2020b] analisou o conteúdo de feedbacks de um curso a distância com base nos níveis de feedback propostos por [Hattie and Timperley 2007]. Baseado nessas boas práticas de feedback, trabalhos recentes propuseram treinar classificadores para identificar aspectos relevantes do feedback [Cavalcanti et al. 2021a]. Por exemplo, o trabalho de [Cavalcanti et al. 2020b] teve como objetivo avaliar classificadores binários para os níveis de feedback propostos por [Hattie and Timperley 2007]. Os autores extraíram 116 recursos do texto usando LIWC, Coh-metrix e recursos adicionais, como o número de Entidades Nomeadas, presença de elogios, presença de saudação e polaridade de sentimento. O trabalho também analisa as características de feedback mais influentes que predizem a qualidade do feedback. Contudo, os trabalho encontrados na literatura sobre esse tema utilizam algoritmos tradicionais de aprendizagem de máquina para lidar com um problema binário. Essa abordagem pode gerar limitações em relação ao desempenho já que neste caso as categorias são completamente isoladas [Kowsari et al. 2019]. Além disto, não encontramos trabalhos utilizando aprendizado profundo para identificação de qualidade de feedback.

Nesse contexto, este artigo propõe uma nova abordagem que aplica aprendizagem profundo para classificar automaticamente mensagens de feedback de professores com base nos níveis de feedback de [Hattie and Timperley 2007]. Mais especificamente, este trabalho apresenta uma abordagem de classificação multi-classe utilizando o BERT (acrônimo de *Bidirectional Encoder Representations from Transformers*) e uma rede BiLSTM (acrônimo de *Bidirectional Long Short-Term Memory*) para classificar as mensagens de feedback. Além disto, o problema de classificação de feedback vem sendo tradicionalmente tratado como categorização binária. Este artigo apresenta pela primeira vez uma alternativa fazendo classificação multi-rótulo, que alcançou resultados melhores que os trabalhos anteriores.

2. Trabalhos relacionados

Para melhorar a qualidade do feedback fornecido aos alunos, alguns trabalhos na literatura propuseram modelos ou princípios que ajudam a aumentar o impacto do feedback na aprendizagem do aluno. Por exemplo, no trabalho de [Hattie and Timperley 2007], um modelo é proposto para a construção de feedback efetivo. Este modelo identifica três perguntas principais que o feedback eficaz deve responder: *"Para onde vou?"*, *"Como estou indo?"*, *"Para onde ir depois?"*. Cada pergunta de feedback opera em quatro níveis: feedback da tarefa (FT), feedback sobre o processamento da tarefa (FP), feedback sobre a auto-regulação (FR) e feedback sobre a pessoa (FS).

O nível a que pertence o feedback influencia na sua eficácia, razão pela qual o feedback centrado nas qualidades do trabalho realizado e no processo ou nas estratégias utilizadas dá maior ajuda ao aluno. O feedback que orienta o aluno para o desenvolvimento de estratégias de autorregulação também tende a ser mais eficaz. Contudo, os comentários focados nas características pessoais dos alunos geralmente são muito vagos e não levam o aluno a focar em seu aprendizado [Brookhart 2017].

[Osakwe et al. 2022] explora o uso da análise de conteúdo automatizada para examinar o feedback fornecido pelos instrutores de acordo com os níveis de feedback propostos por Hattie e Timperley. Os autores usaram o classificador XGBoost (*eXtreme Gradient Boosting*) e um conjunto de dados com textos de feedback escritos em inglês. Os resultados indicam um desempenho de classificação eficaz nos níveis de pessoa, tarefa e processamento da tarefa com valores de precisão de 0.87, 0.82 e 0.69, respectivamente.

O trabalho de [Ruiz Alonso et al. 2022] propõe uma abordagem para a classificação de feedback de acordo com o modelo de feedback de Hattie e Timperley, incorporando uma etapa de ajuste de hiperparâmetros. Os autores realizam experimentos usando os algoritmos SVM (Máquinas de Vetores Suporte), floresta aleatória e k-vizinhos mais próximos. Os autores utilizam textos de feedback gerados por um professor às atividades enviadas pelos alunos em cursos online na plataforma *Blackboard* nos níveis de tarefa, processo, regulação e pessoal propostos no modelo de [Hattie and Timperley 2007]. Os autores obtiveram o melhor resultado usando o classificador SVM com ajuste de hiperparâmetros, alcançando 0,87 de medida F1 e 0,72 de acurácia.

[Cavalcanti et al. 2020a] propôs uma análise de conteúdo do texto de feedback fornecido pelos instrutores com base em princípios de boas práticas de feedback do modelo de [Nicol and Macfarlane-Dick 2006]. Os autores se concentraram em analisar a qualidade do feedback extraído das avaliações coletadas em um curso online oferecido em uma instituição de ensino superior brasileira usando o algoritmo florestas aleatórias. Os autores obtiveram 0,91 de acurácia e 0,82 de *Cohen's kappa* usando um classificador binário que verificava se o texto possuía pelo menos uma boa prática ou nenhuma delas. Na mesma direção, [Cavalcanti et al. 2021b] realizou uma análise semelhante usando o algoritmo XGBoost em conjunto com recursos linguísticos que extraem diversas características dos textos. Com uma abordagem diferente da anterior, os autores consideraram cada boa prática como sendo uma classe binária e alcançaram acurácias de 0,89, 0,77 e 0,89 para os níveis FT, FP e FS, respectivamente.

Não foi encontrado nos trabalhos anteriores abordagens que utilizaram aprendizado profundo na extração de características de mensagens de feedback. Algoritmos de

aprendizado profundo ganharam notoriedade ao apresentarem excelentes resultados em tarefas de visão computacional [Russakovsky et al. 2015] e também em tarefas de Processamento de Linguagem Natural (PLN) [Young et al. 2018]. Esses algoritmos tentam simular o mesmo processo de aprendizado do cérebro humano usando um grande número de conexões geradas em redes neurais profundas. Um exemplo de algoritmo de aprendizado profundo muito utilizado recentemente é o BERT, que é um modelo de linguagem bidirecional treinado em conjuntos de dados muito grandes com base em representações contextuais [Devlin et al. 2018]. O modelo BERT pode ser ajustado usando uma camada de rede neural densa para diferentes tarefas de classificação. Alguns trabalhos já aplicaram algoritmos como BERT no contexto educacional.

Por exemplo, [Junior and Fileto 2021] investigou o uso do BERT para classificar e medir coerência em textos retirados de um fórum educacional com dúvidas de estudantes em disciplinas no Ambiente Virtual de Aprendizagem (AVA) de uma universidade brasileira. Os autores mostraram que o BERT suporta discriminação da ordem de sentenças com até 99,20% de acurácia e também suportou o cálculo de medidas de (in)coerência que permitem discriminar a ordem de sentenças.

O trabalho de [André et al. 2021] avaliou o desempenho de algoritmos baseados em florestas aleatórias e o BERT para detecção automática de Presença Social em discussões online. Os autores comparam a abordagem com mineração de texto tradicional e recursos linguísticos como LIWC e Coh-metrix com a abordagem usando o modelo de linguagem BERT ajustado para classificação de presença social.

Dessa forma, este trabalho difere dos trabalhos mencionados com base em duas contribuições principais: (1) Utilizar um algoritmo de aprendizado profundo para classificação automática de feedback com base no modelo proposto em [Hattie and Timperley 2007]; (2) Avaliar a classificação multi-classe para os níveis FT, FP e FS [Hattie and Timperley 2007].

3. Método

3.1. Dados

O conjunto de dados usado neste trabalho é o mesmo usado por [Cavalcanti et al. 2020a] e foi gerado a partir de um ambiente virtual de aprendizado (AVA) usado em cursos online em uma universidade pública no Brasil. O conjunto de dados contém feedback individual fornecido por instrutores através da ferramenta de envio de atividades no AVA. O conjunto de dados é composto por 1.000 exemplos de mensagens de feedback. Cada feedback foi classificado por especialistas que analisaram o texto do feedback e determinaram se ele pertencia aos 4 níveis propostos por [Hattie and Timperley 2007]). Ou seja, o texto de feedback recebeu o rótulo 0 se não pertencesse a nenhum nível ou o rótulo 1 se pertencesse a algum nível. A Tabela 1 mostra a divisão da base de dados, em termos de níveis, por classe.

3.2. Preparação dos dados

A fase de preparação compreende as etapas de limpeza dos dados e de pré-processamento. Elas tem como objetivo preparar o texto para que ele possa ser usado como entrada para modelos de aprendizado de máquina. Neste trabalho a preparação dos dados foi realizada através da extração de linhas vazias e com valores nulos, remoção de pontuações,

Tabela 1. Divisão da base de dados para os níveis de feedback.

	FT	FP	FR	FS
Classe 0	112	499	992	849
Classe 1	888	501	8	151
Total	1000	1000	1000	1000

números, caracteres únicos isolados, espaços repetidos, *stopwords*, verificação ortográfica e tokenização. Ao final desses processamentos, foram eliminados 4 documentos que continham apenas valores nulos. Essa fase é necessária para remover possíveis ruídos nos dados que podem diminuir a efetividade dos modelos [Bagla et al. 2021].

Além dos processos de limpeza de dados listados acima foi realizada uma diminuição do vocabulário utilizado. O tamanho das sentenças no conjunto de treino varia entre 20 e 30 tokens. Com base nessa informação e nas limitações de *hardware*, optou-se por limitar o tamanho das sentenças de entrada para 30 tokens. Nesse sentido, foi criado um vocabulário de 5000 palavras, baseando no conjunto de treino, onde cada palavra foi associada a um número inteiro. Em seguida, os textos de ambos os conjuntos, treino e teste, foram transformados em sequências numéricas, ou, mais especificamente, vetores de inteiros. Finalmente, para que o requisito de 30 tokens por sentença fosse atendido, foi feito o *padding* de todas essas sentenças para que se encaixassem nesse requisito.

3.3. Extração das características

Após finalização da fase de preparação dos dados foi realizada a extração das características utilizadas na predição dos níveis de feedback. O objetivo desta etapa é transformar o texto em vetores de características numéricas que podem ser processados por algoritmos de aprendizado de máquina, ou de aprendizado profundo, porém, conservando o sentido original do texto. Para isso, foram utilizados modelos de representação de linguagem pré-treinados. Eles são usados para criar representações vetoriais textuais onde palavras com o mesmo significado possuem representações similares. Para este trabalho, foram utilizados diferentes modelos do estado da arte de *word embeddings* [Wang et al. 2020]: Word2vec, GloVe, FastText e o BERT.

Os modelos do Word2vec, GloVe e FastText empregados neste estudo foram treinados em um corpus multi-gênero em língua portuguesa [Hartmann et al. 2017]. Para o experimentos realizados neste estudo foram empregadas os vetores com 50 dimensões. Por outro lado, o modelo BERT aqui utilizado foi o BERTimbau[Souza et al. 2020]. Ele foi treinado no brWaC ([Wagner Filho et al. 2018]), um corpus brasileiro de páginas *web* que contém 2,68 bilhões de *tokens* extraídos de 3,53 milhões de documentos. A versão desse modelo usada neste artigo foi o *base* que gera um vetor de 768 dimensões. Estes modelos foram utilizados por serem os que alcançaram melhores resultados em trabalhos anteriores.

3.4. Seleção de modelos

Para que se pudesse comparar de forma mais precisa o desempenho do método proposto com os trabalhos relacionados que trataram o mesmo problema [Cavalcanti et al. 2021b], foi seguido o mesmo procedimento de particionamento de dados. Foram alocados, portanto, 70% e 30% das amostras para os conjuntos de treino e teste, respectivamente. Pela

mesma razão, na execução dos experimentos, levou-se em consideração apenas as classes FT (devolutiva sobre a tarefa), FP (devolutiva sobre o processo) e FS (devolutiva sobre a pessoa), excluindo a classe FR por ter pouca quantidade de exemplos positivos na base de dados estudada. A Tabela 2 apresenta a distribuição das instâncias de treino e teste por classe analisada.

		Classe positiva	Classe negativa	Total
FT	**Treino**	622 (98,24%)	75 (10,76%)	697
	Teste	265 (88,62%)	34 (11,28%)	299
FP	**Treino**	357 (51,22%)	340 (48,78%)	697
	Teste	144 (48,16%)	155 (51,84%)	299
FS	**Treino**	599 (85,94%)	98 (14,06%)	697
	Teste	53 (17,72%)	246 (82,28%)	299

Tabela 2. Distribuição dos dados após particionamento

Como já foi mencionado nas seções anteriores, este trabalho avalia a utilização de classificadores binários e multi-rótulo para identificação automática das classes FT, FP e FS. Para a classificação binária foi utilizado o *embeddings* do BERT, para a qual foram gerados três modelos, um para cada classe. Por outro lado, na classificação multi-rótulo foi avaliado quatro modelos distintos, um para cada tipo de modelos de representação de linguagem utilizados.

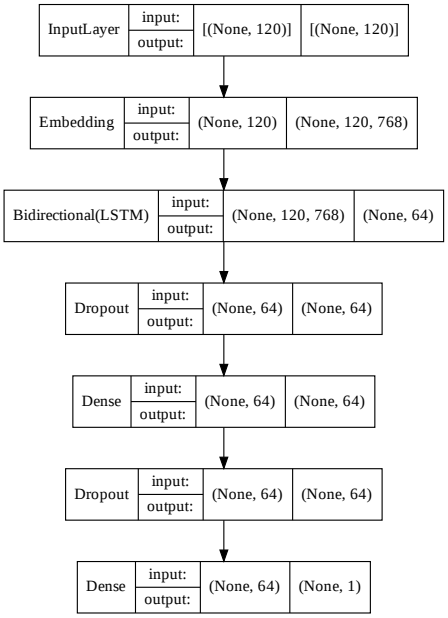

Figura 1. Camadas da rede neural profunda de classificação binária.

No primeiro caso, para classificação binária, o modelo começa com uma camada de entrada que recebe um vetor de 30 inteiros. Essa sequência é então passada para camada de *embedding* onde cada valor inteiro no vetor é substituído por sua representação correspondente. Os *embeddings* são passados para a camada de BiLSTM composta por

duas LSTMs com 32 células de memória. O uso de uma camada de BiLSTM foi embasado em trabalhos anteriores onde para tarefas de classificação elas apresentaram um desempenho superior a seus pares tanto no quesito métricas, quanto no quesito tempo de treino [Graves and Schmidhuber 2005]. Após a BiLSTM, segue-se uma camada de *dropout* cujo a função é aleatoriamente retirar neurônios da rede neural durante o treino e o seu principal objetivo é diminuir a chance do modelo computacional gerado sofrer com *overffiting* [Young et al. 2018]. Logo em seguida, dá-se uma camada densa com 64 neurônios e a função de ativação ReLu introduzida [Xu et al. 2015]. À essa camada, segue-se mais uma camada de *dropout* e, finalmente, uma camada densa com 1 neurônio e a função de ativação *sigmoid* [Young et al. 2018]. Essa função gera um valor real entre 0 e 1 que, para o problema tratado neste artigo, é traduzido como uma indicação da probabilidade de um *feedback* pertencer àquela classe.

No segundo caso, a arquitetura utilizada é bastante semelhante à primeira. Contudo, para o problema de classificação multi-rótulo, foi necessário realizar uma mudança na camada de saída da rede. Embora a função de ativação continue sendo a *sigmoid*, o número de neurônios é de três. Essa camada gera um vetor com três valores reais entre 0 e 1, que, no problema discutido neste artigo, significa a probabilidade de um *feedback* pertencer a uma determinada classe. Além disso, como se trata de um problema de classificação multi-rótulo, o texto de entrada pode pertencer às três classes ao mesmo tempo. Isso acontece justamente por causa da função *sigmoid*. Se no lugar dela fosse usada a função de ativação *softmax* teríamos, ao invés disso, uma rede de classificação multi-classe. As Figuras 1 e 2 apresentam a disposição das camadas em cada uma das arquiteturas propostas.

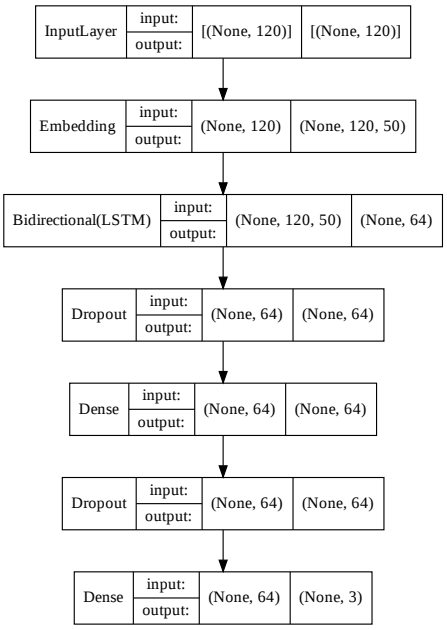

Figura 2. Camadas da rede neural profunda usada para o GloVe, Word2vec e FastText.

3.5. Avaliação dos resultados

Para avaliar o desempenho do classificador binário foram empregadas as seguintes medidas: acurácia e *Cohen's kappa*. A acurácia, segundo [Hossin and Sulaiman 2015], mensura a razão entre as classificações positivas e negativas corretas e o total de amostras avaliadas. O *Cohen's kappa*, por sua vez, é uma métrica estatística que mede a concordância interna entre anotadores em um problema de classificação [Cohen 1960]. As fórmulas 1 e 2 mostram como foi calculada a acurácia e o *Cohen's kappa*, respectivamente. Foram utilizadas essas medidas para realizar uma comparação direta com os trabalhos anteriores.

$$\textit{Acurácia} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\textit{Cohen's kappa} = \frac{p_{\mathrm{o}} - p_{\mathrm{e}}}{1 - p_{\mathrm{e}}} \tag{2}$$

Para o classificador multi-rótulo foram utilizadas as medidas de precisão, cobertura, F_1 e acurácia, que são as medidas mais utilizadas nesse tipo de classificador [Kowsari et al. 2019]. A precisão é uma métrica estatística usada para indicar a porcentagem de amostras corretamente classificadas como positivas considerando todas as que foram assim classificadas. A cobertura é mostra a porcentagem das amostras positivas classificadas corretamente. Por fim, a pontuação $_1$ é a média harmônica entre essas duas últimas métricas, passando uma ideia geral do desempenho do modelo. As fórmulas 3 e 4 mostram como foram calculadas a precisão, cobertura e pontuação F_1, respectivamente.

$$\textit{Precisão} = \frac{TP}{TP + FP} \qquad cobertura = \frac{TP}{TP + FN} \tag{3}$$

$$\textit{F-measure} = \frac{2 \times \textit{precisão} \times cobertura}{\textit{precisão} + cobertura} \tag{4}$$

4. Resultados

4.1. Classificação binária

O objetivo da primeira parte dos experimentos é comparar o desempenho de uma abordagem usando redes neurais profundas e um modelo de representação linguística (BERT) com uma usando o algoritmo XGBoost e indicadores linguísticos extraídos pelo Coh-Metrix e pelo LIWC, em suas versões para português brasileiro proposta por [Cavalcanti et al. 2021b]. Na Tabela 3, é possível observar que nossa abordagem usando BERT + BiLSTM conseguiu obter melhores resultados para os níveis FT e FS comparado aos resultados obtidos por [Cavalcanti et al. 2021b]. Foram obtidos os valores de 0,92 de acurácia e 0,53 de Kappa para o nível FT, 0,75 de acurácia e 0,50 de Kappa para o nível FP e 0,95 de acurácia e 0,80 de Kappa para o nível FS.

Nossa abordagem obteve aumentos de 3,37% na acurácia e 26,19% no Kappa para o nível FT e 6,7% na acurácia e 42,85% no Kappa para o nível FS. Entretanto, nossa abordagem teve resultados que ficaram um pouco abaixo comparado aos resultados usando o XGBoost para o nível FP, com diferença de 2,6% na acurácia e 10% no Kappa. Esses resultados mostram que o uso do BERT + BiLSTM trouxe um avanço

	BERT + BiLSTM		XGBoost	
	Acurácia	**Kappa**	**Acurácia**	**Kappa**
FT	**0,92**	**0,53**	0,89	0,42
FP	0,75	0,50	**0,77**	**0,55**
FS	**0,95**	**0,80**	0,89	0,56

Tabela 3. Análise comparativa entre as pontuações obtidas pela abordagem proposta por [Cavalcanti et al. 2021b] e as deste trabalho.

promissor na classificação de feedback para os níveis de tarefa e pessoa do modelo de [Hattie and Timperley 2007].

4.2. Classificação multi-rótulo

A segunda etapa dos experimentos teve o objetivo de avaliar a abordagem classificação multi-rótulo. Nesse sentindo, a análise foi feita levando em consideração o Word2vec, GloVe, FastText e o BERT e o seu desempenho para cada uma das três classes do conjunto.

A Tabela 4 apresenta os resultados da classe FT. Os resultados mostram uma eficiência e igualdade entre os resultados de todos os modelos, com uma leve melhora do BERT and relação a precisão. Por outro lado, para a FP, resultados apresentados na Tabela 5, embora o BERT ainda apresente a maior precisão e F_1, o Word2vec foi o que obteve melhor cobertura. Finalmente, para a classe FS (Tabela 6), o Word2vec atingiu a melhor pontuação em precisão, mas a maior cobertura e pontuação F_1 foi obtida pelo BERT. Se apenas a F_1 fosse levada em consideração, o BERT seria o modelo de representação de linguagem com melhor performance.

	Word2vec	**Glove**	**FastText**	**BERT**
Precisão	0.89	0.89	0.89	**0.90**
Cobertura	1.00	1.00	1.00	1.00
Pontuação F_1	0.94	0.94	0.94	0.94

Tabela 4. Pontuações das redes neurais profundas para a classe FT.

	Word2vec	**GloVe**	**FastText**	**BERT**
Precisão	0.73	0.70	0.69	**0.79**
Cobertura	**0.75**	0.72	0.67	0.74
Pontuação F_1	0.74	0.71	0.68	**0.76**

Tabela 5. Pontuações das redes neurais profundas para a classe FP.

	Word2vec	**GloVe**	**FastText**	**BERT**
Precisão	**1.00**	0.80	0.86	0.88
Cobertura	0.19	0.45	0.47	**0.57**
Pontuação F_1	0.32	0.58	0.61	**0.69**

Tabela 6. Pontuações das redes neurais profundas para a classe FS

4.3. Discussões

Este trabalho tem como principais contribuições a proposta da utilização de aprendizado profundo e uma abordagem de classificação multi-rótulo para o problema de identificação de qualidade de feedback. Os resultados apresentados mostram que a abordagem aqui proposta alcançou melhores resultados que os artigos anteriores [Cavalcanti et al. 2020b, Cavalcanti et al. 2021b]. Além disto, é importante destacar que um grave problema das abordagens anteriores é o fato da pouca generalização para classes muito desbalanceadas [Osakwe et al. 2022, Barbosa et al. 2020], o modelo de aprendizado profundo proposto neste artigo conseguiu lidar com esse problema sem necessitar de nenhum algoritmo de balanceamento. Além disto, a abordagem multi-rótulo, ainda não utilizada para o problema de análise de feedback, apresentou resultados promissores nas diversas configurações avaliadas.

É importante destacar que os classificadores propostos neste trabalho irão servir como base para o desenvolvimento de uma ferramenta de suporte ao docente na criação de feedback. A classificação nos níveis de feedback propostos por [Hattie and Timperley 2007] pode ser usada num mecanismo de recomendação de boas práticas para professores criarem melhores devolutivas. Além disto, a abordagem proposta também pode ser facilmente adaptada para outros critérios de qualidade de feedback (como as boas práticas propostas por [Nicol and Macfarlane-Dick 2006]) já que estes também são um conjunto de classes binárias de uma mesma mensagem [Cavalcanti et al. 2020a].

5. Limitações e trabalhos futuros

Apesar dos resultados relevantes quando comparado com a literatura, é importante destacar algumas limitações do trabalho: (i) a base de dados é desbalanceada em relação as classes FT e FS, além de não haver instâncias necessárias para classificação da FR. Isso pode afetar a generalização dos resultados encontrados. Além disso, os dados são compostos por mensagens de feedback da mesma instituição de ensino. Como trabalhos futuros, sugere-se a avaliação dos algoritmos multi-rótulos em outros tipos de dados, não apenas mensagens de feedback, assim como a avaliação em dados de outras instituições ou, até mesmo, outros idiomas. Além disto, os modelos desenvolvidos devem ser incorporados na plataforma Tutoria[1] para auxiliar professores a enviarem mensagem de feedback de qualidade para os alunos.

Referências

André, M., Mello, R. F., Nascimento, A., Lins, R. D., and Gašević, D. (2021). Toward automatic classification of online discussion messages for social presence. *IEEE Transactions on Learning Technologies*, 14(6):802–816.

Bagla, K., Kumar, A., Gupta, S., and Gupta, A. (2021). Noisy text data: Achilles' heel of popular transformer based nlp models. *arXiv preprint arXiv:2110.03353*.

Barbosa, G., Camelo, R., Cavalcanti, A. P., Miranda, P., Mello, R. F., Kovanović, V., and Gašević, D. (2020). Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 605–614.

[1]https://dev.tutor-ia.com/

Boud, D. and Falchikov, N. (2007). *Rethinking assessment in higher education: Learning for the longer term*. Routledge.

Brookhart, S. M. (2017). *How to give effective feedback to your students*. ASCD.

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., and Mello, R. F. (2021a). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027.

Cavalcanti, A. P., de Mello, R. F. L., de Miranda, P. B. C., and de Freitas, F. L. G. (2020a). Análise automática de feedback em ambientes de aprendizagem online. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 892–901. SBC.

Cavalcanti, A. P., Diego, A., Mello, R. F., Mangaroska, K., Nascimento, A., Freitas, F., and Gaševic, D. (2020b). How good is my feedback? a content analysis of written feedback. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge - LAK*. ACM.

Cavalcanti, A. P., Mello, R. F., Miranda, P., Nascimento, A., and Freitas, F. (2021b). Utilização de recursos linguísticos para classificação automática de mensagens de feedback. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 861–872. SBC.

Coates, H., James, R., and Baldwin, G. (2005). A critical examination of the effects of learning management systems on university teaching and learning. *Tertiary education and management*, 11:19–36.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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*.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610. IJCNN 2005.

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131, Porto Alegre, RS, Brasil. SBC.

Hattie, J. and Gan, M. (2011). Instruction based on feedback. In *Handbook of research on learning and instruction*, pages 263–285. Routledge.

Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1):81–112.

Henderson, M., Ajjawi, R., Boud, D., and Molloy, E., editors (2019). *The Impact of Feedback in Higher Education: Improving Assessment Outcomes for Learners*. Springer International Publishing, Cham, Switzerland. Google-Books-ID: WyxQxgEACAAJ.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Joulani, P., Gyorgy, A., and Szepesvári, C. (2013). Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461.

Junior, O. O. B. and Fileto, R. (2021). Investigando coerência em postagens de um fórum de dúvidas em ambiente virtual de aprendizagem com o bert. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 749–759. SBC.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., and Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4):150.

Laurillard, D. (1993). *Rethinking University Teaching: Rethinking University Teaching: a Framework for the Effective Use of Educational Technology*. Routledge.

Nicol, D. J. and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218.

Osakwe, I., Chen, G., Whitelock-Wainwright, A., Gašević, D., Cavalcanti, A. P., and Mello, R. F. (2022). Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence*, 3:100059.

Ruiz Alonso, D., Zepeda Cortés, C., Castillo Zacatelco, H., and Carballido Carranza, J. L. (2022). Hyperparameter tuning for multi-label classification of feedbacks in online courses. *Journal of Intelligent & Fuzzy Systems*, (Preprint):1–9.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, 18(2):119–144.

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)*.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wang, S., Zhou, W., and Jiang, C. (2020). A survey of word embeddings based on deep learning. *Computing*, 102(3):717–740.

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network.

Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75.

Ypsilandis, G. (2002). Feedback in distance education. *Computer Assisted Language Learning*, 15(2):167–181.