Aplicação de verbos como proxy para identificação automática do nível cognitivo de questões: uma abordagem baseada na taxonomia de Bloom

Daniyel N. N. Rocha, Cláudio E. C. Campelo, Caio L. M. Jerônimo

Departamento de Sistemas e Computação – Universidade Federal de Campina Grande (UFCG) – Campina Grande – PB – Brasil

{daniyel,caiolibanio}@copin.ufcg.edu.br, campelo@dsc.ufcg.edu.br

Abstract. *The use of questions provides an important mechanism for promoting the development of learning. Thus, educational items can be characterized through methods such as the cognitive levels of Bloom's taxonomy and sets of verbs associated with each one of them. Therefore, this work proposes a new approach for the automatic classification of questions. This was done by building classifiers that were trained with features built with lexicons based on the taxonomy's action verbs, in contrast to existing solutions. The feasibility of this solution was demonstrated, with the results indicating an average F1 of 0.51 and 0.55 for the different versions of lexicons produced.*

Resumo. *A utilização de questões fornece um importante mecanismo para promover o desenvolvimento da aprendizagem. Por sua vez, itens educacionais podem ser caracterizados através de métodos como os níveis cognitivos da taxonomia de Bloom e conjuntos de verbos associados a cada um deles. Diante disto, este trabalho propõe uma nova abordagem para a classificação automática de questões. Isso foi feito através da construção de classificadores treinados com features construídas com léxicos baseados nos verbos de ação da taxonomia, em contraste com as soluções existentes. Foi demonstrado a viabilidade desta solução, com os resultados indicando um F1 médio de 0,51 e 0,55 para as diferentes versões de léxicos produzidas.*

1. Introdução

No âmbito educacional existem diversos instrumentos que são utilizados para promover o desenvolvimento de estudantes e estimular a aquisição de conhecimento. Dentro deste processo de ensino, a aplicação de questões é um dos recursos mais utilizados, permitindo avaliar a aprendizagem em relação ao tema lecionado e fomentar a aquisição de conhecimento. Além disso, a literatura demonstra diversos benefícios com relação ao uso de questões na área educacional, incluindo abordagens direcionadas ao ensino de computação: facilitam a assimilação de novas informações por meio de perguntas que estimulam os alunos a se auto explicarem [Chi et al. 1994]; viabilizam aos estudantes de Ciência da Computação uma melhor compreensão do próprio conhecimento [Tenenberg and Murphy 2005]; permitem melhorar a decomposição de problemas computacionais [Lane and VanLehn 2005]; e incentivam estudantes a construírem conceitos e estratégias, aumentando a motivação e o desenvolvimento cognitivo [Yu et al. 2005].

DOI: 10.5753/sbie.2021.218656

Na maioria das vezes, o desenvolvimento de questões educacionais é feito por profissionais qualificados, como professores e instrutores. Por meio deles, as avaliações são elaboradas e podem reunir diversas abordagens produzidas pelas pesquisas acadêmicas, conforme foi mencionado anteriormente. Entretanto, a construção de exames contendo questões que incorporam estes diferentes aspectos pode evidenciar alguns cenários que exigem atenção. A elaboração manual de itens para construção de avaliações ou banco de questões possui um alto custo associado. Além disso, podem ser obtidos conjuntos de questões desbalanceados em relação à complexidade, contendo, por exemplo, uma maioria de itens que trabalham apenas com fatores superficiais do aprendizado. Ademais, diante do último cenário apresentado, estudos sugerem que avaliações balanceadas são mais eficazes em melhorar o aprendizado dos estudantes [Swart 2009].

A taxonomia de Bloom [Bloom et al. 1956] é um sistema bastante conhecido que propõe uma série de objetivos educacionais para favorecer o processo de aprendizagem e auxiliar educadores e estudantes. Em particular, este modelo apresenta a ideia de um domínio cognitivo fundamentado em diferentes níveis hierárquicos, onde cada um deles se baseia em verbos que trabalham com diferentes aspectos do aprendizado, fornecendo direcionamentos para a criação e avaliação de itens educacionais. Além disso, a taxonomia original foi atualizada ao longo dos anos, se tornando mais representativa e efetiva - estimulando o desenvolvimento de diferentes formas de pensamento. Considerando a versão revisada da taxonomia [LW et al. 2001], os Níveis Cognitivos da Taxonomia de Bloom (NCTB) seguem uma ordem onde os primeiros níveis abordam elementos cognitivos mais simples e os últimos níveis abrangem habilidades mais complexas para o aprendizado, sendo descritas por: Relembrar (*Remembering*), Compreender (*Understanding*), Aplicar (*Applying*), Analisar (*Analyzing*), Avaliar (*Evaluating*) e Criar (*Creating*). Ademais, por se tratar de uma estrutura hierárquica, as categorias correspondem a graus distintos de profundidade das etapas do ensino, permitindo a adoção deste critério para a construção de avaliações [Galhardi and Azevedo 2013] e classificação da complexidade de questões [Padó 2017].

Conforme observado, a taxonomia de Bloom fornece uma estrutura apropriada para a confecção e categorização de itens educacionais. Além disto, este sistema tem como um dos seus elementos centrais os verbos associados aos diferentes NCTB, também conhecidos como verbos de ação. Desta forma, questões educacionais podem ser construídas incorporando estes termos, conferindo uma forma para caracterizar as habilidades que o item procura desenvolver. Além disso, utilizar os verbos para elaborar perguntas pode melhorar a identificação dos objetivos e resultados esperados, pois fornecem mais clareza ao que é desejado e facilitam a avaliação do aprendizado.

O problema de classificar automaticamente questões segundo os NCTB tem recebido atenção de diversos trabalhos e diferentes abordagens foram aplicadas para lidar com esta tarefa. [Jayakodi et al. 2015, Omar et al. 2012] construíram classificadores baseados em regras, onde o primeiro reportou uma acurácia de 82% e realizou uma análise de similaridade textual para estruturar um conjunto de regras, ao passo que o segundo utilizou especialistas para analisar questões e construir regras a partir de padrões gramaticais de cada NCTB. No entanto, estes trabalhos foram avaliados com *datasets* contendo questões de um único domínio e com poucas perguntas - 62 e 30, respectivamente - podendo ocasionar problemas como *overfitting*. Em contrapartida, alguns trabalhos propuseram abor-

dagens baseadas no algoritmo Support Vector Machine (SVM) para empregar modelos de AM [Kusuma et al. 2015, Yahya and Osman 2011]. Entretanto, o primeiro deles fez uso de um conjunto com apenas 130 questões, unicamente no idioma indonésio, enquanto o segundo - mesmo reportando acurácia aproximada de 87% - apresenta resultados de *Recall* (29%) e *F-Score* (39%) não tão altos quanto a métrica anterior, com os autores atribuindo este resultado em razão do pequeno tamanho do *dataset* utilizado - contendo 272 questões no total. Por fim, [Mohammed and Omar 2020] apresentaram bons resultados (*F-score* de até 89%) em abordagens baseadas em AM, onde foram exploradas diferentes *features* sintáticas e semânticas, assim como diferentes *datasets*. Entretanto, não foram avaliados importantes algoritmos de AM (tais como os baseados em árvore de decisão), bem como não foram exploradas *features* baseadas em léxicos, como os verbos de ação da taxonomia de Bloom - instrumentos frequentemente indicados para a classificação de itens e objetivos educacionais [Diab and Sartawi 2017].

Diante disso, este trabalho propõe uma nova abordagem para classificar automaticamente questões de acordo com os NCTB, produzindo modelos de classificação através de *features* baseadas em conjuntos de léxicos, construídos com os verbos de ação. O primeiro grupo de léxicos foi definido a partir de uma coleção dos verbos mais comuns em cada nível da taxonomia [Stanny 2016], enquanto que o segundo conjunto é uma extensão do anterior, mas emprega um método de *Data Augmentation* (DA) baseado em *word embeddings* [Fast et al. 2016] para gerar termos relacionados à cada NCTB. Desta maneira, foram treinados diferentes classificadores fazendo uso das *features* baseadas em léxicos, gerando 6 modelos distintos que foram testados e avaliados. Os resultados indicam que é possível utilizar os verbos de ação para a classificação automática de questões, segundo os NCTB. Além disso, esta nova solução, em conjunto com técnicas de DA, produz resultados ainda melhores - estimulando a utilização de ambientes que empregam os verbos de ação para classificar perguntas educacionais.

Este artigo está organizado da seguinte forma: a Seção 2 define as questões de pesquisa e detalha a metodologia adotada; a Seção 3 apresenta e discute os resultados obtidos e, por fim, a Seção 4 expõe as conclusões e considerações finais.

2. Metodologia

Nesta seção, apresentamos a metodologia utilizada para condução deste trabalho, fornecendo mais detalhes sobre as questões de pesquisa, a composição da base de dados, os modelos de AM, a construção de *features* e as técnicas de PLN que foram aplicadas.

2.1. Questões de pesquisa e hipóteses

Neste trabalho, desenvolvemos uma nova abordagem para classificação automática de perguntas de acordo com os NCTB, baseando-se em léxicos associados aos verbos de ação da taxonomia de Bloom. A condução deste estudo se fundamentou na seguinte questão de pesquisa (QP): *"É possível construir modelos de AM utilizando os verbos de ação da taxonomia de Bloom para definir features e classificar automaticamente questões de acordo com os NCTB?"*.

De modo a responder a QP, a hipótese (H1) a ser trabalhada neste artigo foi definida como: *"É possível implementar modelos de AM utilizando features baseadas nos verbos de ação da taxonomia de Bloom para classificar automaticamente questões de acordo com os NCTB?"*.

As hipóteses descritas anteriormente foram testadas por meio de métricas, que permitiram responder o questionamento suscitado pela QP. As métricas utilizadas para avaliação foram *Precision* (P), *Recall* (R) e *F1-Score* (F1). Estas medidas são frequentemente utilizadas para avaliação de modelos de AM, sendo úteis para analisar a qualidade dos resultados produzidos pelos modelos de classificação.

2.2. Dados

O *dataset* utilizado deriva de três conjuntos de dados distintos [Mohammed and Omar 2020, Yahya et al. 2012, Li and Roth 2002], sendo composto por 2025 questões abertas de língua inglesa e com todas rotuladas manualmente segundo os NCTB. A partir dessas bases de dados, foi feito um tratamento para remoção de duplicatas e elas foram combinadas, constituindo o *dataset* que é base deste estudo - a Tabela 1 apresenta uma amostra de perguntas contidas neste conjunto de dados gerado.

Tabela (1) Exemplos de questões segundo os níveis cognitivos

NCTB	Questão
Remembering	Defina o mercantilismo.
Understanding	Explique nas próprias palavras como criar uma consulta em um banco de dados.
Applying	Aplique sua compreensão do espírito olímpico para desenvolver um novo lema ou slogan.
Analyzing	Compare dois comerciais de alimentos para cães. Qual é a diferença entre eles e como eles tanto vender seus produtos?
Evaluating	Você pode defender a ideia de que o incidente de Simon com a cabeça do porco é o mais místico na história?
Creating	Você pode inventar um outro personagem para a história?

A Figura 1 demonstra a distribuição das perguntas do *dataset* de acordo com os NCTB. Através dela verifica-se um desbalanceamento nos dados, com a maioria das questões sendo do nível *Remembering*, enquanto que os níveis cognitivos *Understanding*, *Applying* e *Analyzing* possuem uma distribuição mais parecida. Por outro lado, os níveis mais altos, *Creating* e *Evaluating*, estão bem menos presentes no *dataset* - ressaltando que perguntas com maior complexidade (segundo a taxonomia) são menos exploradas por conjuntos de dados reais.

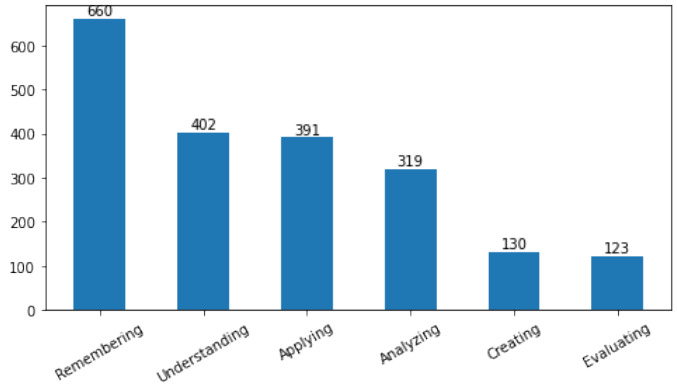


Figura (1) Distribuição dos dados de acordo com os níveis cognitivos

Diante do exposto, o conjunto construído foi separado em subconjuntos de treino e teste, seguindo a proporção de 80% e 20%, respectivamente. Desta forma, o conjunto de treino dispõe de 1620 questões e foi utilizado em etapas posteriores - para a etapa de incremento dos dados (Seção 2.4), extração de *features* (Seção 2.5) e treino dos modelos (Seção 2.6). Além disso, o conjunto de teste foi exclusivamente utilizado na Seção 3, onde os resultados foram obtidos.

2.2.1. Verbos de ação da taxonomia

A partir de uma coleção de verbos propostos por [Stanny 2016] contendo os verbos comuns para cada um dos NCTB, também denominados como verbos de ação. Este acervo tem como objetivo orientar professores e instrutores com relação a quais termos associados aos NCTB são mais significativos, viabilizando a elaboração de questões e exames com mais qualidade. Além disto, é demonstrado que a inserção apropriada dos verbos facilita a avaliação do aprendizado, já que eles permitem descrever melhor os itens educacionais.

Este conjunto utilizado foi definido através da seleção de várias listagens disponíveis no *Google* contendo os verbos mais frequentemente associados à cada NCTB. Desta forma, os termos coletados tiveram as variações gramaticais padronizados e foram removidos registros duplicados em um mesmo NCTB. Além disto, foram atribuídos valores de frequência para cada verbo, correspondendo a quantidade de vezes que um termo foi atribuído a um determinado nível da taxonomia. Por fim, foi aplicado um filtro de qualidade considerando a concordância entre diferentes avaliadores. Ao final, um total de 176 verbos únicos foi obtido, onde cada um contém uma frequência associada e está designada a um NCTB.

Desta maneira, o presente trabalho fez uso da coleção de verbos proposta como forma de caracterizar os diferentes níveis da taxonomia de Bloom. Além disto, esta compilação serviu como base para a definição de um conjunto de léxicos que foi aproveitado em diversas etapas deste trabalho. Ele foi utilizado para a construção de um vocabulário de termos (relativos à taxonomia) e as frequências associadas foram aplicadas à construção de novas *features*. Exemplos de verbos para cada nível podem ser descritos por: *listar* e *identificar* (*Remembering*); *explicar* e *descrever* (*Understanding*); *ilustrar* e *demonstrar* (*Applying*); *analisar* e *categorizar* (*Analyzing*); *comparar* e *julgar* (*Evaluating*); *construir* e *planejar* (*Creating*).

2.3. Pré-processamento dos dados

Esta etapa é fundamental em trabalhos de PLN pois procura remover dados que podem ter algum efeito negativo para o aprendizado dos modelos de classificação. Desta forma, o pré processamento utilizado neste trabalho considerou apenas o conteúdo textual das perguntas, removendo números, pontuações e transformando o todo texto para letra minúscula - sendo aplicado aos conjuntos de treino e teste que foram definidos anteriormente. Além disso, foram removidas das questões todas as stop words, que são palavras muito frequentes na língua inglesa, com exceção das *wh-words* e *how*. Essas palavras foram mantidas por serem importantes para etapas subsequentes, como a construção de features. Por fim, realizamos uma etapa conhecida como lematização, onde cada verbo nas sentenças interrogativas foi transformado para sua forma inflexionada, denominado *lemma*. Assim, é possível interpretar as palavras de acordo com seu significado, independente da flexão original do termo.

2.4. Data augmentation

Modelos de AM dependem da qualidade e também do tamanho dos dados utilizados para treiná-los, podendo ter o desempenho prejudicado por esta razão. Para evitar que os modelos de classificação não consigam se ajustar aos dados e generalizar suficientemente

bem, técnicas de DA são comumente aplicadas a esse tipo de cenário. Para este estudo, foi utilizado um conjunto de quatro estratégias de DA disponibilizadas pela ferramenta *Easy Data Augmentation* (EDA) [Wei and Zou 2019]. Entre os mecanismos disponibilizados, foram utilizados a substituição de termos aleatórios por sinônimos e a inserção do sinônimo de uma palavra em uma posição arbitrária. Além disto, também foram aplicadas a permutação de termos e a exclusão de um vocábulo, novamente de forma aleatória.

Utilizando a ferramenta EDA, realizamos a aplicação das técnicas descritas anteriormente para aumentar os dados de treino, obtido conforme especificado na Seção 2.2. Para esta etapa, especificamos a geração de 16 novas questões a partir de uma única questão, resultando num total de 8515 questões. Desta forma, a situação de desbalanceamento entre os níveis que ocorria anteriormente foi corrigida, permitindo que não haja uma classe que prevaleça muito mais do que as outras. Assim, foi possível minimizar possíveis vieses dos modelos e viabilizar a classificação dos NCTB de modo a terem a mesma importância para os classificadores.

2.4.1. Expansão lexical baseada em *word embeddings*

A partir da coleção de léxicos baseados nos verbos de ação (definidos na Seção 2.2.1), a ferramenta Empath [Fast et al. 2016] foi utilizada para expandir os elementos desta coleção. Esta abordagem permite empregar um conjunto de termos como entrada para a ferramenta e, assim, produzir uma categoria contendo novos léxicos semanticamente similares. Isso é possível em razão do Empath utilizar *word embeddings* para representação das palavras, facilitando a captura de características semânticas dos verbos e identificando melhor os termos relacionados.

Desta forma, os verbos de ação relacionados aos NCTB serviram de entrada para o Empath gerar novas versões expandidas dos níveis existentes e, assim, constituir um novo conjunto de léxicos aumentado. Assim, foi possível definir uma variante da coleção de léxicos original que também foi utilizada para a construção de *features*, nas etapas posteriores deste trabalho.

2.5. Feature engineering

Após pré-processar e organizar os dados, foi preciso transformá-los para um formato compreensível aos modelos. Esta etapa é necessária para a construção dos classificadores pois, em resumo, eles são modelos matemáticos e não lidam com dados textuais. Com isso, os conteúdos referentes às perguntas e aos rótulos dos NCTB precisaram ser convertidos para uma representação interpretável aos algoritmos de construção dos modelos, permitindo que eles aprendam padrões dos dados e a classifiquem novas entradas.

Durante esta etapa, foram conduzidos experimentos para identificar como o conteúdo textual poderia ser transformado, viabilizando a criação das *features* que seriam empregues na etapa de treinamento dos modelos de classificação. Com isso, duas estratégias foram colocadas em prática para a construção das *features*: testar diferentes formas de organização dos dados e também fazer uso de métodos conhecidos, como *Term Frequency – Inverse Document Frequency* (TF-IDF) e *Word Mover's Distance* (WMD) [Kusner et al. 2015]. Este trabalho empregou as abordagens pré-definidas mencionadas anteriormente e também fez uso de uma estratégia que considera a frequência do conjunto de léxicos correspondentes aos verbos de ação, definido na Seção 2.2.1. Neste caso, as

features foram construídas tendo o TF-IDF como base, onde as versões que seguem as outras estratégias (WMD e frequência) são concatenadas com este. Por fim, os léxicos definidos nas Seções 2.2.1 (léxico original) e 2.4.1 (léxico aumentado) foram aplicados como vocabulários para a construção de *features* com quantidades variáveis de termos.

2.5.1. Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF é uma estratégia da área de recuperação de informação que procura verificar a importância de um termo dentro de um conjunto de palavras. Isto posto, o TF-IDF também costuma ser utilizado para a extração de *features*, onde pesos são atribuídos às palavras de documentos. Neste caso, termos menos frequentes têm seus pesos ajustados, sendo considerados mais relevantes - enquanto que os mais comuns serão penalizados. Desta forma, o TF-IDF foi utilizado para mapear valores numéricos às palavras ao conteúdo textual das questões. No entanto, para realizar este mapeamento, foram considerados apenas os termos presentes no vocabulário correspondente ao conjunto de léxicos da taxonomia, definido na Seção 2.2.1. Ao final, uma matriz de *features* foi obtida, com cada questão sendo representada por uma linha e as colunas correspondendo aos verbos de ação da taxonomia de Bloom.

2.5.2. Word Mover's Distance

Outra abordagem utilizada pelo presente trabalho para a construção de *features* foi o *Word Mover's Distance* (WMD), tendo como objetivo calcular a similaridade semântica entre diferentes documentos textuais. Este método também é baseado na ideia de *word embeddings* para representação do conteúdo textual, que são construídos por meio de vetores pré-treinados baseados no algoritmo word2vec [Mikolov et al. 2015]. Assim, o WMD faz uso desta representação para computar a distância entre documentos dentro do espaço vetorial - ou seja, quanto menor o valor resultante, mais semanticamente similares vão ser os textos.

De forma mais específica, os *word embeddings* serviram para representar o conteúdo textual das questões e também dos léxicos associados a cada NCTB. Então, o WMD foi aplicado para calcular a distância semântica destes dois elementos (questão e léxicos de um nível). Desta forma, foi obtida uma matriz de *features* onde as linhas correspondem as sentenças e as colunas são representadas pelos conjuntos de termos dos níveis, com os valores correspondendo às distâncias das questões em relação a cada conjunto de léxicos.

2.5.3. Frequência dos léxicos

Finalmente, através da coleção proposta por [Stanny 2016] e definida na Seção 2.2.1, os dados referentes às frequências dos verbos de ação foram aplicados para a construção de *features*. Neste caso, esta característica corresponde à quantidade de vezes que o verbo foi identificado nas listas coletadas.

Para construir a nova *feature*, a abordagem utilizada também definiu uma matriz, como a empregue para o WMD. Mais especificamente, a matriz foi gerada contabilizando a frequência dos termos (presentes nas questões), restringindo estes termos de modo a serem considerados apenas os conjuntos de léxicos, definidos anteriormente. Então, estes valores de frequência foram atualizados para corresponderem às frequências dos verbos

de ação. Assim, a matriz resultante é composta por linhas que se referem as perguntas e as colunas são os termos do conjunto de léxicos, com os valores correspondendo às frequências dos termos em cada nível da taxonomia.

2.6. Treino e avaliação dos modelos

Nesta etapa os modelos foram treinados e avaliados. Para isto, as *features* construídas nas etapas anteriores foram utilizadas por algoritmos selecionados para realizar o treinamento dos classificadores, onde eles se ajustaram aos dados e, assim, permitir a classificação de novos dados. Desta forma, os seguintes algoritmos foram escolhidos para proceder com esta etapa: *Support Vector Machine* (SVM), *Random Forest* (RF) e *Extreme Gradient Boosting* (XGBoost). A opção pelo SVM se deu por ser um método clássico e frequentemente presente nos trabalhos envolvendo classificação de questões. Por outro lado, RF e XGBoost são aplicações de árvores de decisão que empregam técnicas de *ensemble*, com o primeiro sendo baseado em *Bagging* e o segundo em *Boosting*. Assim, a escolha foi feita em razão de serem mais robustos que o SVM [Caruana and Niculescu-Mizil 2006].

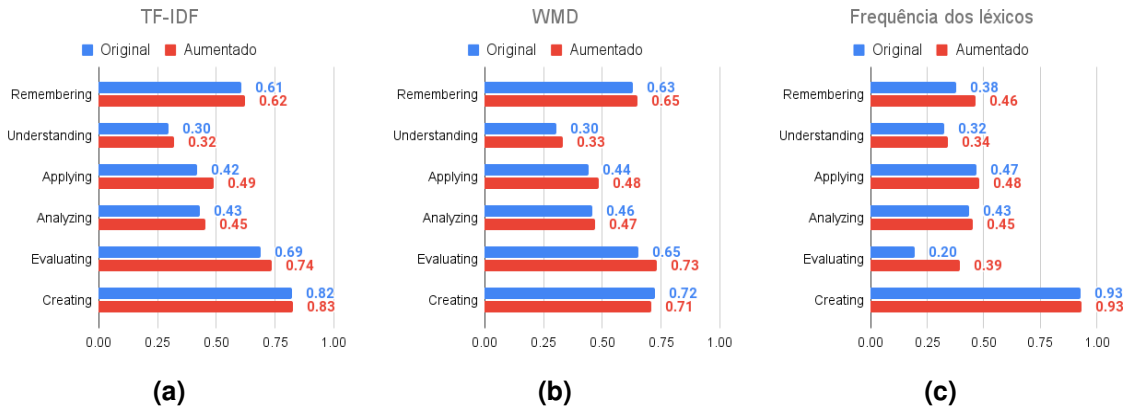
Além disto, os modelos de classificação foram treinados levando em conta hiperparâmetros previamente selecionados. Para a definição destes valores, várias configurações foram testadas em um esquema de validação cruzada, permitindo a obtenção de hiperparâmetros que produzem modelos com performances superiores em relação aos ajustes originais. Além disso, a utilização da validação cruzada permite que os modelos sejam treinados e avaliados considerando todo o conjunto de dados, fornecendo configurações que assimilaram a totalidade dos dados. Com isso, estes modelos testados durante a etapa de seleção dos hiperparâmetros são capazes de se ajustar melhor aos dados empregues no treinamento.

Para este trabalho, a seleção de hiperparâmetros foi feita com um esquema de busca aleatória, onde cada um dos conjuntos de valores compondo um novo modelo que foi testado e avaliado. Em relação aos dados, o conjunto de treino foi utilizado para a validação cruzada com 4 subconjuntos mutuamente exclusivos e estratificados. Assim, para cada algoritmo, obtemos coleções de hiperparâmetros que definiram melhores modelos. A partir disto, os dados de teste foram utilizados para a produção dos resultados, que serão avaliados e discutidos na Seção 3.

3. Resultados

A partir dos algoritmos selecionados, foram treinados vários modelos de classificação, viabilizando a obtenção de respostas para as hipóteses idealizadas pela questão de pesquisa. A Figura 2 sintetiza os resultados dos modelos treinados com as diferentes *features*, em razão do F1-Score médio. Neste caso, estão sendo considerados os classificadores baseados nos léxicos originais e aumentados e os valores segundo os NCTB. Assim, os resultados produzidos pelos modelos utilizaram as *features* construídos com o conjunto de treinamento e foram avaliados com os dados de teste. As Tabelas 2 e 3 fornecem mais detalhes com relação a performance dos modelos. Nas tabelas, é possível observar os desempenhos detalhados dos classificadores, construídos a partir das diferentes *features*, e também o desempenho médio dos modelos, utilizando cada variação de *feature*. Além disto, estes valores correspondem aos múltiplos NCTB existentes.

Com base no comparativo apresentado pela Figura 2 observamos que a utilização do método que aumenta a quantidade de dados e lida com o desbalanceamento oferece



(a) **(b)** **(c)**

Figura (2) Comparativo dos F1-Scores médios por NCTB, considerando as diferentes versões de léxicos e segundo as *features* baseadas em: (a) TF-IDF; (b) WMD; (c) Frequência dos léxicos.

certa melhora nos resultados - apenas no nível *Creating* da *feature* WMD a performance apresentou uma leve piora. Analisando os valores obtidos em razão de cada uma das *features*, verificamos que os modelos baseados em *features* TF-IDF e WMD possuem, de forma geral, resultados bastante semelhantes e superiores aos baseados em frequências. Isso pode ser um indício de que incorporar as frequências dos léxicos aos modelos não resulta em modelos superiores. Por outro lado, as *features* que são construídos considerando a estrutura das palavras em relação às frases (TF-IDF) e aspectos semânticos (WMD) ofereceram resultados mais bem distribuídos para os NCTB.

Tabela (2) Resultados dos modelos treinados com diferentes *features*, baseados no léxico original

NCTB	Modelos	TF-IDF			Frequência dos léxicos			WMD		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Remembering	SVM	0.44	0.95	0.60	0.88	0.24	0.38	0.44	0.95	0.60
	Random Forest	0.44	0.97	0.61	0.86	0.24	0.38	0.51	0.85	0.64
	XGBoost	0.44	0.98	0.61	0.91	0.24	0.37	0.50	0.93	0.65
Understanding	SVM	0.60	0.18	0.28	0.94	0.20	0.32	0.60	0.18	0.28
	Random Forest	0.73	0.20	0.31	0.86	0.22	0.35	0.85	0.21	0.33
	XGBoost	0.67	0.20	0.30	0.88	0.18	0.30	0.67	0.20	0.30
Applying	SVM	0.85	0.29	0.43	1.00	0.31	0.48	0.85	0.29	0.43
	Random Forest	0.81	0.28	0.41	1.00	0.29	0.45	0.72	0.36	0.48
	XGBoost	0.88	0.28	0.42	1.00	0.31	0.48	0.91	0.26	0.41
Analyzing	SVM	1.00	0.26	0.41	0.90	0.28	0.42	1.00	0.26	0.41
	Random Forest	0.95	0.29	0.45	0.95	0.29	0.45	0.62	0.43	0.51
	XGBoost	1.00	0.28	0.43	1.00	0.28	0.43	0.57	0.37	0.45
Evaluating	SVM	0.59	0.81	0.68	0.10	0.94	0.19	0.59	0.81	0.68
	Random Forest	0.60	0.78	0.68	0.11	0.97	0.20	0.49	0.88	0.63
	XGBoost	0.63	0.81	0.71	0.11	1.00	0.20	0.54	0.81	0.65
Creating	SVM	0.77	0.87	0.82	0.95	0.91	0.93	0.77	0.87	0.82
	Random Forest	0.79	0.83	0.81	0.91	0.91	0.91	0.54	0.91	0.68
	XGBoost	0.80	0.87	0.83	1.00	0.91	0.95	0.53	0.91	0.67

A partir dos resultados logrados e considerando a H1 definida na Seção 2.1, podemos avaliar melhor esta hipótese. Com isso, verificou-se que foi possível implementar modelos de classificação a partir de métodos de AM baseados nos verbos de ação dos níveis cognitivos da taxonomia de Bloom. Levando em conta que a média das métricas foi P = 0,72, R = 0,54 e F1 = 0,51 (*features* utilizando léxicos originais) e P = 0,72, R = 0,58 e F1 = 0,55 (*features* com léxicos aumentados), observamos que os resultados são

satisfatórios e, mesmo com valores não muito altos, esta nova abordagem indica a viabilidade da classificação automática de questões segundo os diferentes NCTB, bem como a utilização de *features* baseadas em verbos de ação da taxonomia de Bloom.

Tabela (3) Resultados dos modelos treinados com diferentes *features*, baseados no léxico aumentado

NCTB	Modelos	TF-IDF			Frequência dos léxicos			WMD		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Remembering	SVM	0.47	0.93	0.62	0.89	0.25	0.39	0.47	0.92	0.62
	Random Forest	0.47	0.97	0.63	0.86	0.24	0.38	0.53	0.90	0.66
	XGBoost	0.46	0.98	0.62	0.45	1.00	0.62	0.52	0.93	0.67
Understanding	SVM	0.59	0.20	0.29	1.00	0.18	0.31	0.59	0.20	0.29
	Random Forest	0.70	0.23	0.35	0.91	0.24	0.38	0.83	0.24	0.38
	XGBoost	0.71	0.21	0.32	0.94	0.21	0.34	0.64	0.22	0.33
Applying	SVM	0.80	0.35	0.49	1.00	0.31	0.48	0.80	0.35	0.49
	Random Forest	0.88	0.36	0.51	0.89	0.31	0.46	0.88	0.35	0.50
	XGBoost	0.93	0.31	0.47	1.00	0.34	0.50	0.96	0.30	0.46
Analyzing	SVM	0.80	0.31	0.44	0.76	0.29	0.42	0.80	0.31	0.44
	Random Forest	0.79	0.34	0.47	0.78	0.32	0.46	0.53	0.51	0.52
	XGBoost	0.83	0.31	0.45	0.84	0.32	0.47	0.62	0.35	0.45
Evaluating	SVM	0.60	0.88	0.71	0.11	0.94	0.19	0.56	0.88	0.68
	Random Forest	0.66	0.84	0.74	0.12	0.97	0.21	0.66	0.91	0.76
	XGBoost	0.67	0.88	0.76	0.69	0.91	0.78	0.67	0.88	0.76
Creating	SVM	0.76	0.83	0.79	0.95	0.87	0.91	0.76	0.83	0.79
	Random Forest	0.87	0.87	0.87	0.92	0.96	0.94	0.69	0.96	0.80
	XGBoost	0.77	0.87	0.82	1.00	0.91	0.95	0.38	0.91	0.54

4. Conclusão

A aplicação de questões educacionais provê inúmeros benefícios para o processo de ensino e aprendizagem, sendo um dos métodos mais utilizados dentro da educação. No entanto, a adoção de critérios para caracterização destes itens se faz importante, auxiliando instrutores e viabilizando a construção de avaliações balanceadas e mais eficazes.

Este trabalho propôs a utilização dos níveis cognitivos da versão revisada da taxonomia de Bloom e os verbos de ação associados como uma forma de representação de perguntas e construção de *features*. Desta maneira, foi experimentada uma abordagem para o treinamento de modelos de classificação por meio de léxicos fundamentados nestes verbos. Além disto, modelos baseados em árvores de decisão foram selecionados e avaliados, realizando-se um tratamento nos dados para evitar problemas decorrentes de um *dataset* com poucos dados. Desta forma, os classificadores construídos apresentaram um F1 aproximado de 0,51 e 0,55 para os conjuntos de léxicos desenvolvidos, demonstrando um resultado regular, bem como a viabilidade deste método.

A classificação automática de questões educacionais permite que bancos de questões sejam construídos e a elaboração de exames mais efetivos em avaliar os estudantes. Como forma de continuar este trabalho, pretendemos utilizar esta abordagem de léxicos com outros métodos de construção de modelos, como os baseados em redes neurais e aprendizagem profunda. Além disso, é possível produzir novas *features* combinando os atributos utilizados por este artigo para, assim, construir novos modelos. Desta maneira, poderemos representar melhor as questões e incrementar a capacidade preditiva com outras abordagens. Assim, esta solução pode ser aplicada como método para auxiliar professores no desenvolvimento de questões educacionais mais bem fundamentadas e facilitar a avaliação do aprendizado, por se basearem na taxonomia de Bloom.

Referências

Bloom, B. S., Engelhart, M. B., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (1956). *Taxonomy of educational objectives. The classification of educational goals. Handbook 1: Cognitive domain*. Longmans Green, New York.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

Chi, M. T., De Leeuw, N., Chiu, M.-H., and LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477.

Diab, S. and Sartawi, B. (2017). Classification of questions and learning outcome statements (LOS) into blooms taxonomy (BT) by similarity measurements towards extracting of learning outcome from learning material. *CoRR*, abs/1706.03191.

Fast, E., Chen, B., and Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657.

Galhardi, A. C. and Azevedo, M. M. d. (2013). Avaliações de aprendizagem: o uso da taxonomia de bloom. In *Anais do VII Workshop Pós-graduação e Pesquisa do Centro Paula Souza, São Paulo*, volume 1, pages 237–247.

Jayakodi, K., Bandara, M., and Perera, I. (2015). An automatic classifier for exam questions in engineering: A process for bloom's taxonomy. In *2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pages 195–202. IEEE.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Kusuma, S. F., Siahaan, D., and Yuhana, U. L. (2015). Automatic indonesia's questions classification based on bloom's taxonomy using natural language processing a preliminary study. In *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 1–6. IEEE.

Lane, H. C. and VanLehn, K. (2005). Teaching the tacit knowledge of programming to noviceswith natural language tutoring. *Computer Science Education*, 15(3):183–201.

Li, X. and Roth, D. (2002). Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

LW, A., DR, K., PW, A., KA, C., Mayer, R., PR, P., Raths, J., and MC, W. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. A. (2015). Computing numeric representations of words in a high-dimensional space. US Patent 9,037,464.

Mohammed, M. and Omar, N. (2020). Question classification based on bloom's taxonomy cognitive domain using modified tf-idf and word2vec. *PloS one*, 15(3):e0230442.

Omar, N., Haris, S. S., Hassan, R., Arshad, H., Rahmat, M., Zainal, N. F. A., and Zulkifli, R. (2012). Automated analysis of exam questions according to bloom's taxonomy. *Procedia-Social and Behavioral Sciences*, 59:297–303.

Padó, U. (2017). Question difficulty–how to estimate without norming, how to use for automated grading. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*, pages 1–10.

Stanny, C. J. (2016). Reevaluating bloom's taxonomy: What measurable verbs can and cannot say about student learning. *Education Sciences*, 6(4):37.

Swart, A. J. (2009). Evaluation of final examination papers in engineering: A case study using bloom's taxonomy. *IEEE Transactions on Education*, 53(2):257–264.

Tenenberg, J. and Murphy, L. (2005). Knowing what i know: An investigation of undergraduate knowledge and self-knowledge of data structures. *Computer Science Education*, 15(4):297–315.

Wei, J. and Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6383–6389.

Yahya, A. A. and Osman, A. (2011). Automatic classification of questions into bloom's cognitive levels using support vector machines. In *The International Arab Conference on Information Technology, Naif Arab University for Security Science (NAUSS)*, pages 1–6.

Yahya, A. A., Toukal, Z., and Osman, A. (2012). Bloom's taxonomy–based classification for item bank questions using support vector machines. In *Modern advances in intelligent systems and tools*, pages 135–140. Springer.

Yu, F.-Y., Liu, Y.-H., and Chan, T.-W. (2005). A web-based learning system for question-posing and peer assessment. *Innovations in education and teaching international*, 42(4):337–348.