Agrupamento automático de mensagens em fóruns educacionais

Fábio Mariano[1], **Valmir Macário**[1], **Rafael Ferreira Mello**[1,2]

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

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

fabiio.mariiano@gmail.com , {valmir.macario,rafael.mello}@ufrpe.br

Abstract. *The internet brought numerous advantages when it comes to facilitating access to information. However, a common problem that makes it difficult for teachers to monitor is information overload. In order to mitigate this, this article performs groupings using the K-Means, K-Medoids and Agglomerative algorithms in 1652 posts from 4 different educational forums of a higher education course in order to group similar messages to help the teacher, having to deal with with a smaller amount of information. In each post, it extracts features and applies NLP techniques, in addition to using a vector representation for the text of the posts. Finally, it evaluates the quality of each cluster using the following metrics: Silhouette and Davies-Boulding coefficients.*

Resumo. *A internet trouxe inúmeras vantagens quando a questão é facilitar o acesso a informação. Porém, um problema comum que dificulta o acompanhamento dos professores é a sobrecarga de informações. Com intuito de mitigar isto, este artigo realiza agrupamentos utilizando os algoritmos K-Means, K-Medoids e o Aglomerativo em 1652 postagens de 4 fóruns educacionais diferentes de um curso superior agrupando as mensagens semelhantes para auxiliar o professor, lidando com uma quantidade menor de informação. Em cada postagem, extrai características e aplica técnicas de PLN, além de utilizar uma representação vetorial para o texto das postagens. Por fim, avalia a qualidade dos agrupamentos utilizando as métricas: silhueta e Davies-Boulding.*

1. Introdução

O aprendizado online, do inglês *online learning* [Anderson 2009], tem raízes na tradição da Educação a Distância (EAD), o qual iniciou-se há vários anos atrás com os primeiros cursos por correspondência. Com o advento da internet, o potencial de alcançar alunos de todo o mundo aumentou muito. A EAD se beneficia diretamente disso, conectando alunos e professores que estão há milhares de quilômetros de distância, além de contar com inúmeras outras vantagens como: flexibilidade de tempo e/ou redução de custos [Morilhas 2009]. Dessa forma, o ensino a distância vem se tornando cada vez mais popular pois provê maior flexibilidade para acessar os conteúdos e instruções a qualquer hora, de qualquer lugar [Means et al. 2009].

Há diferentes tipos de plataformas para auxiliar o aprendizado online chamadas de Ambientes Virtuais de Aprendizagem (AVA), do inglês *Virtual Learning Enviroment*, as quais possuem o objetivo de conectar os alunos e os professores, gerando e trocando

DOI: 10.5753/sbie.2022.224635

informações entre eles. Informações essas que podem ser trocadas de forma síncrona ou assíncrona, através de diferentes recursos como fóruns, *chats*, wiki, entre outros. Os recursos que interagem de forma assíncrona tendem a obter dos alunos discursos mais inerentes e auto-reflexivos e, portanto, são mais propícios ao aprendizado profundo do que os recursos que interagem de forma síncrona [Nason 2006]. Dentre esses recursos assíncronos, o fórum é um dos que permite a maior interatividade entre os alunos e os professores [Caspi et al. 2003,Rolim et al. 2020], além de oferecer diversas possibilidades para os professores interagirem com a turma, de forma efetiva [Wever et al. 2006].

Um dos problemas da utilização de AVAs, principalmente utilizando recursos como fóruns, é a sobrecarga de informações [Wulf et al. 2014]. Essas plataformas envolvem centenas ou milhares de estudantes que interagem entre si e com os professores, gerando assim um enorme volume de dados estruturados e, principalmente, não estruturados [Ferreira-Mello et al. 2019, André et al. 2021]. Quanto mais alunos uma turma possuir, mais complicado fica para o professor acompanhar e, além de outras coisas, fornecer feedbacks para eles. Entretanto, o feedback é um dos recursos mais poderosos durante o aprendizado de um aluno, especialmente em cursos a distância, ajudando-o a identificar falhas e a melhorar sua estratégia de aprendizado [Pinheiro et al. 2019, Cavalcanti et al. 2021]. Desta forma, tendo em vista auxiliar os professores no momento de gerar os feedbacks para os alunos e reduzir a quantidade de informação necessária a ser analisada. Uma alternativa é utilizar métodos automáticos para fornecer informações simples e representativas sobre o conteúdo das mensagens, de uma maneira que os ajudem a identificar sua relevância e contexto [Gerosa et al. 2001].

A Mineração de texto pode ser usada com objetivo de mitigar a sobrecarga de informações em fóruns, utilizando-a para extrair informações importantes do texto não estruturado [Berry 2003]. Isto é, os professores e alunos podem se beneficiar de diferentes técnicas da mineração de texto, como Processamento de Linguagem Natural (PLN), classificação e agrupamento de textos, recuperação de informações e resumo de documentos, para extrair informações e conhecimentos interessantes e não triviais de textos não estruturados [Ferreira-Mello et al. 2019].

Uma vez que estamos tratando particularmente dos fóruns nos AVAs que, basicamente, são compostos por conversas entre alunos e professores, a utilização do PLN pode ser bastante benéfica para extração de informação relevante. O PLN é a área responsável pela manipulação de textos ou falas [Chowdhury 2003], utilizando diferentes algoritmos para analise semântica e sintática, em que a maioria das aplicações nas plataformas educacionais são relacionadas a avaliação automática de redações e perguntas discursivas [Ferreira Mello et al. 2022, Passero et al. 2019, Mello et al. 2022].

Outras duas técnicas que podem ser utilizadas no contexto de fóruns educacionais é a classificação, ou aprendizado supervisionado, e o Agrupamento, ou aprendizado não supervisionado, ambas técnicas de aprendizagem de máquina. A classificação categoriza documentos considerando suas características levando em consideração categorias pré-definidas, e o agrupamento categoriza documentos baseados na similaridade entre eles [Hassani et al. 2020]. Ambas as técnicas podem ser utilizadas para diversos objetivos, como análise de sentimentos, classificação de questões, categorização de fóruns, identificar padrões de aprendizagens, entre outros.

Considerando o contexto apresentado, este trabalho propõe a utilização de técnicas de agrupamento para criar grupos de postagens similares, para que o professor consiga direcionar mensagens de resposta (feedback) de forma a agregar mais dúvidas e auxiliar mais alunos. Para este estudo, foi utilizado uma base de dados com 1652 postagens de alunos e professores em um fórum educacional de uma instituição de ensino superior em Pernambuco. A abordagem de agrupamento utiliza técnicas de pré-processamento nos textos para criar uma representação vetorial para os mesmos e informações extraídas relacionadas as postagens (por exemplo a profundidade da postagem em relação à temática), com intuito de melhorar o agrupamento final. Por fim, é realizado o agrupamento de combinações de entradas com os algoritmos *K-Means*, *K-Medoids* e o Hierárquico Aglomerativo. Foram utilizadas medidas clássicas de avaliação de agrupamento, com intuito de analisar a qualidade de cada agrupamento para cada conjunto de entrada.

2. Trabalhos relacionados

Com o crescimento da utilização de AVAs e devido a maioria das interações nestes ambientes feitas pelos estudantes serem por texto, é preciso lidar com um dos maiores problemas desses tipos de ambientes: o problema da sobrecarga de informações [Wulf et al. 2014]. Diante disso, é essencial utilizar métodos automáticos para extrair informações relevantes desses conteúdos, com principal objetivo de auxiliar os professores. Uma possibilidade para lidar com esse problema é a utilização do algoritmos de agrupamento, que agrupa dados em *clusters* de forma que os elementos que o componham sejam mais semelhantes entre si do que com os outros *clusters* [Han et al. 2012]. Neste contexto, vários trabalhos tem aplicado agrupamento para resolver problemas em análise textual.

Balabantaray *et al* [Balabantaray et al. 2015] realizou uma análise com dois algoritmos de agrupamento, o *K-Means* e o *K-Medoids*, com o objetivo de identificar categorias de elementos textuais. Os autores utilizaram uma base de dados com 100 documentos, distribuídos em 20 de cada tipo a seguir: literatura, entretenimento, esportes, política e zoologia . Utilizando o valor de $K = 5$ e as distâncias de Manhattan e Euclidiana, foi concluído que o *K-Means*, utilizando a distância de Manhattan, obteve melhores resultados, agrupando mais documentos semelhantes no mesmo conjunto.

Singh *et al* [Singh et al. 2011] fez uma análise mais completa de diferentes configurações para agrupamento textual. Eles aplicaram os algoritmos *K-Means*, *heuristic K-means* e *fuzzy C-Means* combinando com diferentes representações (*Term Frequency* (TF), *Term Frequency-Inverse Document Frequency* (TF-IDF) e *Boolean*) e técnicas de pré-processamento (com e sem remoção de *stop word* e *stemming*) em uma base de dados composta por artigos de notícias de 3 bases de dados diferentes (Reuters-21578, Classic 4 e 20 Newsgroups), totalizando cerca de 5000 artigos com 59 classes distintas . E, através de experimentos, concluiu que o TF-IDF obteve melhores agrupamentos com todos os algoritmos entre os conjuntos de dados, assim como a utilização do *stemming*. Enquanto aos algoritmos de agrupamento, o *fuzzy C-Means* alcançou melhores resultados.

No contexto educacional, também foram encontrados trabalhos que utilizaram algoritmos de agrupamento para auxiliar os professores. Por exemplo, Pardo *et al* adotou uma abordagem utilizando agrupamento e *learning analytics* [Pardo et al. 2019]. Neste estudo, ao invés do professor gerar um feedback para cada aluno, dependendo de seu desempenho nas atividades da disciplina no AVA do curso, ele criou um conjunto de

mensagens para cada atividade, variando de acordo com o nível de engajamento que o aluno poderá ter nela. Com isso, para cada aluno é gerado um feedback personalizado automático, através do seus traços digitais extraídos do AVA em questão. Dessa forma, foi concluído que obteve-se uma mudança por parte dos alunos na percepção de feedbacks, além de uma pequena a média melhora no desempenho deles a médio prazo.

Lim *et al* [Lim et al. 2019] analisou o impacto de um sistema de feedback baseado em *learning analytics*, aprendizagem autorregulada e o desempenho acadêmico dos estudantes do primeiro ano de graduação do curso de Ciências Biológicas, por três anos. Criando em cada turma dois grupos de tamanhos iguais, em que um dos grupos recebia os feedbacks e o outro não. E concluiu, entre outras coisas, que a nota final do curso foi consideravelmente diferente entre os dois grupos, com o grupo de estudantes que recebia os feedbacks alcançando notas mais altas do que o grupo que não recebia.

Esses dois últimos trabalhos reforçam a importância do feedback para os alunos, o qual têm papel importante de ajudá-los a encontrarem falhas em seus aprendizados e melhorarem seus desempenhos [Pinheiro et al. 2019].

Outra aplicação na área de educação de algoritmos de agrupamento é a análise de fóruns educacionais. Lopez *et al* [Lopez et al. 2012] utilizou a base de dados de um AVA de uma universidade e propôs classificações via abordagem de agrupamento, com a principal finalidade sendo determinar se a participação do estudante no fórum do curso é capaz de predizer a nota final dele . Para isso, compara a acurácia desta abordagem com a acurácia da classificação dos dados utilizando algoritmos clássicos. Este trabalho utiliza algoritmos como: *Random Forest* e *Simple Naive Bayes* para a classificação, e *Simple K-Means*, *Hierarchical Clusterer* e *Expectation Maximisation* para o agrupamento, entre outros. Dos 114 alunos que compõem a base, extraiu 11 atributos de cada, entre eles: número de mensagens enviadas, número de mensagens lidas, tempo gasto no fórum e a nota final obtida. Através de várias combinações, mostrou que o algoritmo *Expectation Maximisation* obteve os melhores resultados.

O trabalho realizado por Ramos *et al* [Ramos et al. 2016] utiliza a base de dados do AVA de um curso com 200 alunos de uma outra universidade e compara os agrupamentos de algoritmos hierárquicos e não hierárquicos, mostrando que ambos os métodos apresentaram resultados semelhantes, formando grupos de tamanhos, dados e características similares.

Por fim, Ferreira Mello *et al* [Ferreira-Mello et al. 2019] realizou uma revisão sistemática da literatura dos últimos 10 anos que envolvessem técnicas de mineração de texto na educação, avaliando exatos 343 artigos. Nesta revisão, os autores indicaram que nenhum deles possuíam objetivo de agrupar o conteúdo de mensagens em fóruns educacionais, sendo isso considerado uma lacuna na área.

Diante do que foi exposto nesta seção, a principal contribuição deste trabalho é a análise de diferentes configurações de algoritmos e características para realizar o agrupamento de postagens em fóruns educacionais com o intuito de auxiliar professores a entender as principais dúvidas (ou grupos de dúvidas) dos alunos. Vale ressaltar, que não foram encontrados na literatura trabalhos nesta linha, por isso a importância deste estudo.

3. Materiais e métodos

3.1. Base de dados

A base de dados utilizada foi a de uma instituição de ensino superior de Pernambuco, a qual contém postagens de 273 alunos e 1 professor em fóruns com temáticas específicas e descritas a seguir, fazendo com que os alunos desenvolvam raciocínios e discutam entre em si sobre o tema. Com um total de 1652 postagens compostas por 4 fóruns do curso de Psicologia do Desenvolvimento, o qual faz parte do curso de graduação de Psicologia. Todos os fóruns aconteceram simultaneamente, entre agosto e dezembro de 2014. Este curso foi escolhido por conter mais mensagens nos fóruns e com mais palavras por mensagem. Foram eles:

- “1º FÓRUM TEMÁTICO: “Pau que nasce torto morre torto””;
- "2º FÓRUM TEMÁTICO: "O Tamanho é de Uma Criança, Mas o Comportamento é de Um Adulto;
- “3º FÓRUM TEMÁTICO:: “NÃO VOU ME ADAPTAR””;
- "4º FÓRUM TEMÁTICO: A escola que afirma não ter bullying ou não sabe o que é, ou está negando a sua existência.".

A Tabela 1 apresenta as estatísticas de cada fórum. O primeiro fórum teve mais postagens, mas com menos conteúdo. Enquanto os outros três fóruns tiveram número similar de postagens e tamanho do texto.

Tabela 1. Estatísticas dos fóruns utilizados

	Número de postagens			
Fórum	Professor	Aluno	Total	**Média de palavras por postagem**
1º FÓRUM TEMÁTICO	190	443	633	31
2º FÓRUM TEMÁTICO	86	303	389	49
3º FÓRUM TEMÁTICO	52	268	320	61
4º FÓRUM TEMÁTICO	46	264	310	63

Para cada fórum, o professor inicia-o com uma postagem temática inicial, a qual descreve o tema e contém um conteúdo introdutório sobre o assunto. Com isso, cada aluno pode tanto responder a essa postagem, como também responder a postagens de outros alunos, gerando assim várias *threads* de discussões. Além disso, o professor também pode responder a postagens dos alunos, a fim de fomentar a discussão entre eles.

3.2. Características para realizar o agrupamento

3.2.1. Pré-processamento dos textos

O pré-processamento foi realizado nas mensagens dos alunos e do professor, com finalidade de fazer uma limpeza nos textos e remover elementos com pouca importância nas mensagens que pudessem atrapalhar ou contribuir negativamente nos resultados, como pontuações, artigos e etc. Com isso, foram utilizadas técnicas de PLN como:

- Normalização: transformando os caracteres para minúsculos e removendo pontuação;

- Remoção de *stopwords*: removendo palavras com pouco significado para o texto;
- *lemmatization*: transformando as palavras na sua forma primitiva, levando em consideração a palavra anterior para assim manter o contexto da palavra em questão, resultando numa melhor precisão. Por exemplo, um verbo conjugado vai para o infinitivo. Outra técnica semelhante é o *stemming*, que também tem finalidade de reduzir a palavra a sua forma primitiva, porém, sem levar em consideração o contexto dela, o que permite a execução ser mais rápido porém com menos precisão. Logo, utilizamos neste trabalho apenas o *lemmatization*.

Além dessas técnicas, também foi necessário remover *tags* HTML dos textos, visto que são salvas junto com o texto para manter a formatação da mensagem.

3.2.2. Representação vetorial dos textos

Para que seja possível a realização de cálculos em cima de textos, se faz necessário a sua representação vetorial. Então, para representar os textos das postagens, foi utilizada a representação vetorial TF-IDF. Largamente utilizada na literatura, a representação vetorial é criada a partir da presença ou ausência de termos, ou baseada na frequência absoluta ou relativa desses termos, como é o caso do TF-IDF [Aguiar and Prati 2015]. O valor TF-IDF de uma palavra aumenta proporcionalmente à medida que aumenta o número de ocorrências dela em um documento, no entanto, esse valor é equilibrado pela frequência da palavra no corpus. Isso auxilia a distinguir o fato da ocorrência de algumas palavras serem geralmente mais comuns que outras. A seguir, a Equação 1 representa o TF-IDF, onde n representa a quantidade de vezes que o termo t aparece em um documento d contendo m termos. Enquanto df é o número de ocorrências no conjunto de documentos N do termo t.

$$TF(t,d) = \frac{n}{m}$$

$$IDF(t) = \log_2\left(\frac{N}{1+df}\right)$$

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \tag{1}$$

3.2.3. Características adicionais

Além de considerar o texto das postagens na análise de agrupamento, também foi realizada uma extração de características dessas postagens, a fim de realizar comparações do agrupamento considerando apenas o conteúdo das mensagens e também considerando outras características. Para isso, foram extraídas de cada postagem as seguintes características:

1. A profundidade da postagem. Isto é, postagens que são respostas diretas à temática possuem profundidade 1, e caso alguém responda a essas postagens, a sua profundidade será de 2, e assim sucessivamente;

2. Se a postagem é do professor ou de um dos alunos;
3. Se a postagem é temático ou não. Isto é, se é a postagem que o professor faz no início do fórum, com a temática da discussão;
4. Se a postagem é resposta direta a temática ou não;
5. Quantidade de palavras na mensagem da postagem, após o pré-processamento;
6. A data da postagem.

3.3. Algoritmos de agrupamento

Para realizar os agrupamentos, serão utilizados 3 algoritmos típicos e bastante utilizados na área educacional [Ferreira-Mello et al. 2019]. O *K-Means*, *K-Medoids* e o Aglomerativo, descritos com mais detalhes a seguir.

O *K-Means* [MacQueen 1967] é um algoritmo não supervisionado de agrupamento que particiona n elementos entre k grupos, onde cada um desses elementos pertencerá ao grupo mais próximo da média. O objetivo é minimizar a distância média dos documentos dos centros de seus *clusters*, onde esses centros são definidos como a média ou centroide dos documentos em um *cluster*.

O algoritmo *K-Medoids* [Kaufmann and Rousseeuw 1987] é uma variação do *K-Means* que é mais robusta a ruídos e pontos fora da curva no conjunto de entradas. Sua principal diferença é que ao invés de usar o ponto médio como centro do grupo, o centro é o ponto cuja a soma de dissimilaridades para todos os elementos no grupo é mínima.

O algoritmo Hierárquico Aglomerativo [Day and Edelsbrunner 1984] faz parte dos algoritmos de agrupamentos hierárquicos, ou *Hierarchical Clustering*, que consistem em realizar uma série de sucessivos agrupamentos a fim de agregar ou desagregar elementos, construindo assim uma hierarquia de *clusters*. O resultado deste agrupamento é representado formando uma árvore de *clusters*, ou Dendograma, o qual pode ser construído utilizando a estratégia *top-down*, partindo da raiz para as folhas, utilizando o método divisivo. Ou a estratégia *bottom-up*, partindo das folhas em direção as raízes, que, por sua vez, é a estratégia utilizada pelo algoritmo Aglomerativo.

3.4. Métricas de avaliação dos agrupamentos

Para avaliação da qualidade do resultados dos algoritmos de agrupamento foram utilizadas métricas clássicas da literatura. São elas:

1. Coeficiente de Silhueta: O coeficiente de silhueta consiste em avaliar a coesão dos *clusters* representando a distância média de uma amostra i a outros *clusters*. Logo, os coeficientes de silhueta do agrupamento devem calcular a média dos coeficientes de cada amostra, gerando assim um valor entre 1 e -1, sendo 1 o melhor e -1 o pior valor [Guo et al. 2019]. Além disso, valores próximos a 0 indicam sobreposição dos *clusters*. através da proximidade de uma amostra i aos outros pontos do *cluster* o qual eles pertencem, e a proximidade do mesmo ponto aos pontos do *cluster* mais próximo a ele.
2. Davies-Boulding: O índice de Davies-Boulding [Davies and Bouldin 1979] mede a média de similaridade entre cada *cluster* e o mais semelhante [Kovács et al. 2006]. Como os *clusters* precisam ser compactos e separados, quanto menor o índice, melhor o agrupamento.

4. Resultados

Pra todos os algoritmos foram utilizadas 4 combinações diferentes de características divididas em 4 conjuntos descritos abaixo:

- O conjunto 1 consiste somente nas mensagens dos alunos e professores, representadas vetorialmente.
- O conjunto 2 contém as mensagens do conjunto 1, somado as características: quantidade de palavras, profundidade da postagem e se é resposta à temática.
- O conjunto 3 possui as mensagens, as características também presentes no Conjunto 2 e, além delas, também: se é uma postagem do professor e se é uma postagem temática.
- Por fim, o conjunto 4 contem as mensagens e todas as características. Isto é, as citadas no Conjunto 3 mais a data da postagem.

Além disto, para todos os algoritmos, foram avaliados o número de clusters K de 2 à 10. A Tabela 2 contém os resultados das avaliações dos agrupamentos, para cada um dos 4 conjuntos. As métricas Coeficiente de Silhueta e Davies-Boulding estão com as respectivas abreviações SS e DB, respectivamente. Para cada coluna, os melhores valores estão em negrito. E, considerando todos os valores, os melhores resultados para cada métrica estão também sublinhados.

5. Discussão dos Resultados

Como podemos observar na Tabela 2, os melhores agrupamentos foram obtidos, para ambas as métricas, nos agrupamentos para o Conjunto 4. Isto fortalece com a informação de que a data da postagem é uma característica importante para realizar um melhor agrupamento. Pois, ela é a única característica extra que o Conjunto 4 possui além do Conjunto 3. É importante destacar que a utilização apenas das características textuais levou aos piores resultados em todos os casos analisados. Isso mostra que, neste cenário, utilizar apenas o texto não seria o suficiente para realizar o agrupamento com qualidade.

Para todos os conjuntos de entrada, observando os resultados para a métrica de silhueta, os algoritmos *K-Means* e *K-Medoids* concentraram todos os melhores resultados. Além disso, esses algoritmos obtiveram resultados bastante parecidos, diferenciando, na maioria das vezes, de apenas décimos, para as duas métricas.

O algoritmo hierárquico aglomerativo obteve valores abaixo dos outros, aconselhando-se seu uso apenas em caso de que se necessite entender a hierarquia na formação dos agrupamentos.

Do ponto de vista de implicações práticas, o trabalho proposto tem como objetivo auxiliar professores a responder postagens de alunos de forma eficiente já que as perguntas estariam agrupadas. Assim, espera-se diminuir quantidade de informação que o professor precisa processar, e espera-se que a solução aumente a quantidade de alunos com respostas adequadas. Desse modo, mesmo sendo comum na literatura iniciar a variação do valor de K a partir de 2, ou seja, dividindo apenas em 2 grupos, a escolha de uma opção com mais grupos é mais viável. O resultado obtido pelo *K-Means* e *K-Medoid* com 5 grupos para o Conjunto 4 foi o melhor resultado obtido pelo índice DB e um dos melhores com o coeficiente da Silhueta, portanto seria a melhor escolha para agrupar as mensagens dos fóruns analisados.

Tabela 2. Resultados das avaliações dos agrupamentos para cada conjunto

	Conjunto 1		Conjunto 2		Conjunto 3		Conjunto 4	
Agrupamento	SS	DB	SS	DB	SS	DB	SS	DB
K-Means - 2	0.020	2.845	**0.657**	0.566	**0.655**	0.569	0.647	0.471
K-Means - 3	0.029	4.841	0.637	0.528	0.634	0.545	**0.667**	0.423
K-Means - 4	0.025	4.578	0.616	0.515	0.615	0.515	0.650	0.442
K-Means - 5	0.034	4.768	0.600	0.509	0.601	0.503	0.657	**0.409**
K-Means - 6	0.041	5.223	0.575	0.517	0.594	0.491	0.635	0.455
K-Means - 7	0.045	4.423	0.581	0.488	0.580	0.489	0.594	0.528
K-Means - 8	0.048	4.786	0.579	0.496	0.578	0.497	0.602	0.506
K-Means - 9	0.058	4.778	0.578	0.534	0.554	0.506	0.609	0.481
K-Means - 10	**0.061**	5.489	0.519	0.527	0.553	0.519	0.605	0.473
K-Medoids - 2	-0.016	**1.032**	0.614	0.589	0.614	0.589	0.653	0.448
K-Medoids - 3	0.006	1.827	0.583	0.591	0.582	0.591	0.662	0.426
K-Medoids - 4	-0.003	1.330	0.580	0.579	0.579	0.580	0.655	0.415
K-Medoids - 5	0.006	2.080	0.584	0.532	0.583	0.533	0.656	0.411
K-Medoids - 6	-0.013	1.959	0.579	0.522	0.578	0.523	0.616	0.471
K-Medoids - 7	0.007	1.722	0.580	0.491	0.572	0.523	0.629	0.472
K-Medoids - 8	-0.034	1.629	0.575	0.498	0.575	0.497	0.600	0.498
K-Medoids - 9	-0.011	1.928	0.533	0.516	0.573	**0.484**	0.611	0.470
K-Medoids - 10	0.004	1.728	0.573	**0.487**	0.547	0.517	0.594	0.474
AG - 2	0.023	3.028	0.614	0.589	0.614	0.589	0.643	0.482
AG - 3	0.032	2.514	0.617	0.582	0.616	0.582	0.615	0.443
AG - 4	0.042	2.389	0.614	0.527	0.613	0.527	0.622	0.494
AG - 5	0.046	3.727	0.523	0.550	0.528	0.547	0.604	0.444
AG - 6	0.052	4.223	0.532	0.540	0.538	0.538	0.613	0.489
AG - 7	0.051	3.884	0.536	0.498	0.543	0.496	0.615	0.482
AG - 8	0.056	4.276	0.532	0.509	0.536	0.509	0.571	0.506
AG - 9	0.059	3.893	0.535	0.515	0.539	0.515	0.579	0.477
AG - 10	0.059	3.643	0.525	0.523	0.529	0.523	0.599	0.451

6. Conclusão

Com o auxílio dos agrupamentos, permitindo separar as postagens dos alunos em grupos semelhantes, o professor pode conseguir mitigar um pouco o problema da sobrecarga de informações [Wulf et al. 2014], lidando com a informação de uma forma mais sintetizada, permitindo assim otimizar o envio de feedback, o qual tem papel importante para melhorar o desempenho dos alunos [Pinheiro et al. 2019].

Os agrupamentos foram realizados num total de 1652 postagens de 4 fóruns diferentes de um curso de Psicologia do Desenvolvimento, incluindo postagens de alunos e professores. Onde, para cada postagem, foi realizado o pré-processamento dos textos utilizando técnicas de PLN e uma representação vetorial, além de extrair características de cada postagem. Após realizar experimentos para 4 conjuntos de combinações de entrada e 3 algoritmos de agrupamentos diferentes, a qualidade dos agrupamentos foram avaliadas usando 2 métricas clássicas de avaliação de agrupamento. Considerando todos os conjuntos de entrada, o conjunto que possuía as mensagens das postagens junto com

todas as características extraídas, Conjunto 4, foi o que obteve os melhores resultados, para todos os algoritmos de agrupamento analisados. Ele também obteve o valor 0.409 no índice Davies-Boulding com os algoritmos *K-Means* e *K-Medoid* com $k = 5$. Sendo 5 a quantidade de agrupamentos nas mensagens dos fóruns analisados mais viável, para que o professor seja auxiliado e consiga lidar com mais alunos e focando nas perguntas e respostas desses 5 grupos de mensagens.

Para os trabalhos futuros, para que o professor consiga utilizar dos agrupamentos em seu dia a dia com os alunos, seria necessário desenvolver um sistema de agrupamento de postagens onde seria possível inserir as postagens como entrada e, após processamento, obtê-las separadas de acordo com cada grupo identificado. Outra ideia seria experimentar o uso de agrupamentos baseados em *Deep Learning* [Aljalbout et al. 2018] e outros algoritmos de agrupamento como o *fuzzy C-Means*. E, por fim, explorar a seleção de características através de algoritmos específicos para isso.

Referências

Aguiar, R. and Prati, R. (2015). Incorporação de representação vetorial distribuída de palavras e parágrafos na classificação de sms spam.

Aljalbout, E., Golkov, V., Siddiqui, Y., and Cremers, D. (2018). Clustering with deep learning: Taxonomy and new methods. *ArXiv*, abs/1801.07648.

Anderson, T. (2009). *The Theory and Practice of Online Learning*. AU Press, Edmonton, AB, CAN, 2nd edition.

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.

Balabantaray, R. C., Sarma, C., and Jha, M. (2015). Document clustering using k-means and k-medoids. *CoRR*, abs/1502.07938.

Berry, M. W. (2003). *Survey of Text Mining*. Springer-Verlag, Berlin, Heidelberg.

Caspi, A., Gorsky, P., and Chajut, E. (2003). The influence of group size on nonmandatory asynchronous instructional discussion groups. *Internet and Higher Education*, 6(3):227–240.

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

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, 37(1):51–89.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Day, W. and Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1332.

Ferreira Mello, R., Fiorentino, G., Oliveira, H., Miranda, P., Rakovic, M., and Gasevic, D. (2022). Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in portuguese. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 404–414.

Gerosa, M. A., Fuks, H., and De Lucena, C. J. P. (2001). Use of categorization and structuring of messages in order to organize the discussion and reduce information overload in asynchronous textual communication tools. In *Proceedings Seventh International Workshop on Groupware. CRIWG 2001*, pages 136–141.

Guo, H., Ma, J., and Li, Z. (2019). Active semi-supervised k-means clustering based on silhouette coefficient. In Xhafa, F., Patnaik, S., and Tavana, M., editors, *Advances in Intelligent, Interactive Systems and Applications*, pages 202–209, Cham. Springer International Publishing.

Han, J., Kamber, M., and Pei, J. (2012). Data mining concepts and techniques, third edition.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., and Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, 4(1):1.

Kaufmann, L. and Rousseeuw, P. (1987). Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416.

Kovács, F., Legány, C., and Babos, A. (2006). Cluster validity measurement techniques. *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*.

Lim, L.-A., Gentili, S., Pardo, A., Kovanović, V., Whitelock-Wainwright, A., Gašević, D., and Dawson, S. (2019). What changes, and for whom? a study of the impact of learning analytics-based process feedback in a large course. *Learning and Instruction*, page 101202.

Lopez, M., Luna, J. M., Romero, C., and Ventura, S. (2012). Classification via clustering for predicting final marks based on student participation in forums. *Proc. of 5th Int. Conf. on Educational Datamining*, pages 148–151.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297.

Means, B., Toyama, Y., Murphy, R., Bakia, M., and Jones, K. (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. *Centre for Learning Technology*.

Mello, R. F., Neto, R., Fiorentino, G., Alves, G., Arêdes, V., Silva, J. V. G. F., Falcão, T. P., and Gašević, D. (2022). Enhancing instructors' capability to assess open-response using natural language processing and learning analytics. In *European Conference on Technology Enhanced Learning*, pages 102–115. Springer.

Morilhas, L. J. (2009). The expansion of distance learning (dl) in brazilian higher education: Trends for the beginning of the next decade. *Future Studies Research Journal: Trends and Strategies*, 1(1):66–88.

Nason, M. (2006). Learning together online: Research on asynchronous learning networks. *Education and Information Technologies*, 11:191–192.

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., and Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, 50(1):128–138.

Passero, G., Ferreira, R., and Dazzi, R. L. S. (2019). Off-topic essay detection: A comparative study on the portuguese language. *Revista Brasileira de Informática na Educação*, 27(03):177–190.

Pinheiro, A., Ferreira, R., Ferreira, M. A., Rolim, V., Freitas, F., and Gasevic, D. (2019). An analysis of the use of good feedback practices in online learning courses.

Ramos, J., Rodrigues, R., Sedraz, J., Gomes, A., and Silva, R. (2016). A comparative study between clustering methods in educational data mining. *IEEE Latin America Transactions*, 14:3755.

Rolim, V., Mello, R. F., and Lins, R. D. (2020). Análise de discussões em fóruns educacionais usando mineração de texto e análise de grafos. *Sociedade Brasileira de Computação*.

Singh, V. K., Tiwari, N., and Garg, S. (2011). Document clustering using k-means, heuristic k-means and fuzzy c-means. In *2011 International Conference on Computational Intelligence and Communication Networks*, pages 297–301.

Wever, B. D., Schellens, T., Valcke, M., and Keer, H. V. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers Education*, 46(1):6 – 28. Methodological Issues in Researching CSCL.

Wulf, J., Blohm, I., Brenner, W., and Leimeister, J. M. (2014). Massive open online courses. *Business Information Systems Engineering*, 6:111–114.