UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Fabio Mariano Costa Silva Gomes Pereira

Agrupamento automático de mensagens em fóruns educacionais

Recife

2022

Fabio Mariano Costa Silva Gomes Pereira

Agrupamento automático de mensagens em fóruns educacionais

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Rafael Ferreira Leite de Mello

Recife

2022

Fabio Mariano Costa Silva Gomes Pereira

Agrupamento automático de mensagens em fóruns educacionais

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Trabalho aprovado. Recife, 01 de Junho de 2022.

Rafael Ferreira Leite de Mello (Orientador)
UFRPE

Valmir Macário Filho
UFRPE

Recife
2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

P436a Pereira, Fabio Mariano Costa Silva Gomes
Agrupamento automático de mensagens em fóruns educacionais / Fabio Mariano Costa Silva Gomes Pereira. - 2022.
28 f.

Orientador: Rafael Ferreira Leite de Mello.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2022.

1. agrupamento. 2. mineração de texto. 3. fórum. 4. educacional. I. Mello, Rafael Ferreira Leite de, orient. II. Título

CDD 004

Dedico este trabalho à minha avó Cristina, a minha mãe Edilene e ao meu filho Pedro.
Vocês me fazem querer sempre buscar o melhor de mim.

Agradecimentos

Agradeço primeiramente e principalmente à minha família. À minha mãe que sempre batalhou muito para criar, sozinha, eu e minhas irmãs da melhor forma possível, e fez isso com excelência. À minha avó paterna que me ensinou tanto e até hoje vive intensamente em meu coração, ecoando em minha mente seus conselhos e nossas eternas lembranças. À minha madrinha, que se faz presente em toda a minha vida desde minha infância até hoje, me dando todo tipo de apoio e suporte necessário onde quer que eu precise, sem medir esforços para isso. Ao meu pai, que mesmo um pouco distante, sempre foi e será um exemplo para mim. Ao meu filho, que me estimula, mesmo sem saber, a ser uma melhor pessoa a cada dia, e me fez descobrir o quão magnífico é ser pai. À minha namorada e futura esposa, por estar ao meu lado em todos os momentos, acreditando em mim e me fazendo tão feliz e realizado. E a todas as pessoas restantes da minha família que estão sempre comigo, me dando forças e apoio.

Agradeço também ao diretor da época da escola que cursei o ensino fundamental e médio, por sempre acreditar no meu potencial e me ajudar a alcançar meu principal objetivo da época, o de ingressar na universidade.

Agradeço imensamente aos meus amigos que fiz na universidade e que, sem eles, esta jornada teria sido incomparavelmente mais difícil. Aos que não mantenho tanto contato, mas que me ajudaram passiva ou ativamente, um enorme agradecimento. E aos que fazem parte até hoje do meu ciclo próximo de amigos principalmente. A companhia nas inúmeras madrugadas seguidas em claro de Douglas. O companheirismo e trabalhos em dupla, sempre buscando e alcançando à excelência, com Rafael. Os ensinamentos de Anderson, meu ponto de referência técnica desde o primeiro período. O prazer de ajudar Tomás na programação. A dedicação de Caio em ajudar os amigos sob qualquer circunstancia. Agradeço a vocês e a todos os outros não citados, mas que me permitiram viver incontáveis momentos marcantes como estes, os quais foram me moldando e me fazendo evoluir como aluno, profissional e, principalmente, como pessoa. O meu profundo e mais sincero agradecimento a todos os meus amigos que me ajudaram e permitiram ser ajudados por mim. Vocês são demais.

Por fim, mas não menos importante, um agradecimento aos meus docentes e servidores que tanto me ensinaram e me ajudaram durante o curso. Em especial

ao meu orientador, que quando ainda era coordenador, no meu segundo período na faculdade, me estimulou a tomar uma das minhas decisões mais importantes na época e que hoje vejo o quanto foi primordial e me fez seguir caminhos que me permitiram chegar até onde estou hoje. Desde lá, inconscientemente já havia escolhido-o como meu orientador. Meu eterno agradecimento por toda a paciência, calma e empatia que sempre teve comigo.

“O insucesso é apenas uma oportunidade para recomeçar com mais inteligência.”
(Henry Ford)

Resumo

A internet, a adoção cada vez mais presente da Educação a Distância e os Ambientes Virtuais de Aprendizagem trazem inúmeras vantagens quando a questão é facilitar o acesso a informação. Porém, um problema comum que dificulta o acompanhamento dos professores e, sobretudo, o envio de feedback, devido a maior quantidade de alunos por turma, quando comparado com o ensino presencial, é a sobrecarga de informações. Com intuito de mitigar isto, este artigo realiza agrupamentos utilizando os algoritmos K-Means, K-Medoids, DB Scan e o Aglomerativo em 1652 postagens de 4 fóruns educacionais diferentes de um curso superior a fim de agrupar as mensagens semelhantes para auxiliar o professor, tendo que lidar 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 de cada agrupamento utilizando as métrica: coeficiente de silhueta e Davies-Boulding.

Palavras-chave: agrupamento, mineração de texto, fórum, educacional.

Abstract

The internet, the increasingly adoption of Distance Education and Virtual Learning Environments bring numerous advantages when it comes to facilitating access to information. However, a common problem that makes it difficult for teachers to monitor and send feedback, due to the greater number of students per class, when compared to face-to-face teaching, which leads to an information overload. In order to mitigate this, this article performs groupings using the K-Means, K-Medoids, DB Scan 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.

Keywords: clustering, text mining, forum, educational.

Lista de tabelas

Lista de abreviaturas e siglas

EAD	Educação a Distância
AVA	Ambientes Virtuais de Aprendizagem
PLN	Processamento de Linguagem Natural
TF-IDF	*Term Frequency-Inverse Document Frequency*
DB Scan	*Density-Based Spatial Clustering of Applications With Noise*

Sumário

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 centenas de anos atrás com os primeiros cursos por correspondência. Com o advento da internet e da *World Wide Web*, o potencial de alcançar alunos de todo o mundo aumentou muito. O EAD então, por sua vez, 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 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; GORSKY; CHAJUT, 2003; ROLIM; MELLO; LINS, 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). 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 uma das influências mais poderosas durante o aprendizado de um aluno, ajudando-o a identificar falhas e a melhorar sua estratégia de aprendizado (PINHEIRO et al., 2019). 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; Fuks; De Lucena, 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 AVAs podem se beneficiar de diferentes técnicas da mineração de texto, como processamento de linguagem natural, 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, é indispensável a utilização do Processamento de Linguagem Natural (PLN). 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., 2019).

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 (FERREIRA-MELLO et al., 2019). 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.

Levando tudo isso em consideração, este trabalho propõe a utilização de técnicas de agrupamento para criar grupos de postagens, identificando postagens similares, para que o professor consiga direcionar mensagens de resposta de forma a agregar mais dúvidas. 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 tanto técnicas de pré-processamento nos textos para criar uma representação vetorial para os mesmos, quanto informações extraídas relacionadas as postagens em si, com intuito de melhorar o agrupamento final, como por exemplo: a profundidade da postagem em relação à temática. Por fim, é realizado o agrupamento de combinações de entradas com os algoritmos *K-Means*, *K-Medoids*, *DB Scan* e 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; KAMBER; PEI, 2012). Neste contexto, vários trabalhos tem aplicado agrupamento para resolver problemas em análise textual.

Balabantaray *et al* 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 (BALABANTARAY; SARMA; JHA, 2015). 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* 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 (TF, 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 (Singh; Tiwari; Garg, 2011). 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* per-

sonalizado 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* 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 (LIM et al., 2019). 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* 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 (LOPEZ et al., 2012). 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* 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 (RAMOS et al., 2016).

Por fim, Ferreira Mello *et al* 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 (FERREIRA-MELLO et al., 2019).

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, entre outras coisas, postagens dos alunos em fóruns com temáticas específicas e descritas a seguir, fazendo com que os alunos desenvolvam raciocínios e discutam entre em si sobre o tema.

Foram selecionadas então 4 fóruns do curso de Psicologia do Desenvolvimento, totalizando 1652 postagens. 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

Fórum	Número de postagens	Média de palavras por postagem
1º FÓRUM TEMÁTICO	633	31
2º FÓRUM TEMÁTICO	389	49
3º FÓRUM TEMÁTICO	320	61
4º FÓRUM TEMÁTICO	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.

Além de considerar o texto das postagens na análise de agrupamento, também foi realizado também uma extração de características dessas postagens, a fim de realizar comparações do agrupamento considerando apenas o conteúdo das mensagens e considerando também outras características. Logo, 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, em milissegundos.

3.2 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.

3.3 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 *Term Frequency-Inverse Document Frequency* (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; 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 fórmula matemática do TF-IDF, onde **n** representa a quantidade de vezes que o parametro t aparece no parametro d, e **m** o total de termos em d:

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

$$\text{IDF(t)} = \log \frac{N}{1 + df}$$

$$\text{TF - IDF(t,d)} = \text{TF(t,d)} * \text{IDF(t)}$$

3.4 Algoritmos de agrupamento

Para realizar os agrupamentos, serão utilizados 4 algoritmos típicos e bastante utilizados na área educacional (FERREIRA-MELLO et al., 2019). O *K-Means*, *K-Medoids*, *DB Scan* e o Aglomerativo, descritos com mais detalhes a seguir.

3.4.1 *K-Means*

O *K-Means* 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 centróide dos documentos em um *cluster* (HARTIGAN; WONG, 1979).

3.4.2 *K-Medoids*

O algoritmo *K-Medoids* 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 (KAUFMANN; ROUSSEEUW, 1987).

3.4.3 *DB Scan*

Density-Based Spatial Clustering of Applications with Noise, ou apenas *DB Scan*, é um dos algoritmos de agrupamento utilizado, o qual consegue agrupar conjuntos de dados com formato arbitrário e de grande tamanho. Após os dados serem representados vetorialmente, ele os particiona em subgrupos baseados na densidade das regiões, e aumenta ou diminui cada grupo de acordo com uma análise de conectividade entre os dados (ESTER et al., 1996).

3.4.4 Aglomerativo

O algoritmo Aglomerativo 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 (DAY; EDELSBRUNNER, 1984).

3.5 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; MA; LI, 2019). Além disso, valores próximos a 0 indicam sobreposição dos *clusters*. através da proximidade de um ponto 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; Bouldin, 1979) mede a média de similaridade entre cada *cluster* e o mais semelhante (KOVáCS; LEGáNY; BABOS, 2006). Como os *clusters* precisam ser compactos e separados, quanto menor o índice, melhor o agrupamento.

4 Resultados

Cada algoritmo foram utilizadas 4 combinações diferentes de características. Segue o detalhe de cada conjunto explorado:

- 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 os algoritmos *K-Means* e *K-Medoids*, foram utilizados os valores de 6 à 10 para o valor de K na avaliação destes algoritmos. Mesmo sendo comum na literatura iniciar a variação do valor de K a partir de 2, ou seja, dividindo apenas em 2 grupos, como o intuito deste trabalho é separar as postagens em grupos que ajudem o professor, dividir em poucos grupos pode não ser eficiente. Para o algoritmo Aglomerativo, foi utilizado o valor 5 para a quantidade de grupos. E o algoritmo DBScan não precisa deste parâmetro de entrada, pois a quantidade de grupos é determinada dinamicamente durante a execução do agrupamento.

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.

Ao final dos experimentos com o algoritmo *DB Scan*, a quantidade de grupos finais, para cada conjunto de entrada, são, respectivamente: 61, 57, 57 e 4.

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 - 6	0.04	5	0.57	0.52	**0.60**	0.50	**0.63**	0.45
K-Means - 7	0.04	4	**0.58**	0.49	0.58	0.49	0.59	0.53
K-Means - 8	0.05	4	0.58	0.50	0.58	0.50	0.60	0.51
K-Means - 9	0.06	4	0.58	0.53	0.55	0.51	0.61	0.48
K-Means - 10	**0.06**	5	0.52	0.53	0.55	0.52	0.60	0.47
K-Medoids - 6	-0.01	2	0.58	0.52	0.58	0.52	0.62	0.47
K-Medoids - 7	0.01	2	0.58	0.49	0.57	0.52	0.63	0.47
K-Medoids - 8	-0.03	2	0.57	0.50	0.57	0.50	0.60	0.50
K-Medoids - 9	-0.01	2	0.53	0.52	0.57	**0.48**	0.61	0.47
K-Medoids - 10	0.00	2	0.57	**0.48**	0.55	0.52	0.59	0.47
DB	-0.09	**1**	-0.65	1	-0.65	1	0.08	0.58
AG	0.05	3	0.58	0.53	0.58	0.50	0.60	**0.44**

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 permite afirmar que a data da postagem é uma característica importante para realizar um melhor agrupamento. Pois, ela é a única característica a mais 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, com valores máximos sempre próximos ou maiores a 60%, indicando um bom agrupamento. Em contrapartida, para a mesma métrica, o algoritmo *DB Scan*, em 3 dos 4 agrupamentos, obteve valores negativos, o que indicam agrupamentos ruins. E, no único que obteve valor positivo, observa-se que a quantidade de grupos foi mais próxima da quantidade de grupos utilizadas nos outros 3 algoritmos, que alcançaram resultados bem melhores. Indicando a princípio que, neste cenário, agrupar em várias dezenas de grupos não atinge bons resultados.

Os algoritmos *K-Means* e *K-Medoids*, ambos muito semelhantes entre si, alcançaram, além de 80% de todos os 8 melhores resultados, valores sempre bastante parecidos. Diferenciando, na maioria das vezes, de apenas décimos, para as duas métricas.

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.

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 4 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 foi o que obteve os melhores resultados, para as 2 métricas. Atingindo o valor 0.63 no coeficiente de silhueta com o algoritmo *K-Means* com k=6. E o valor 0.44 para o Davies-Boulding com o algoritmo Aglomerativo.

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*. Avaliar os diferentes parâmetros dos algoritmos *DB Scan* e do Aglomerativo, além de diferentes conjuntos de características, como por exemplo, utilizar somente as características extraídas mas sem o texto das postagens. E, por fim, explorar a seleção de características através de algoritmos específicos para isso.

Referências

AGUIAR, R.; PRATI, R. Incorporação de representação vetorial distribuída de palavras e parágrafos na classificação de sms spam. In: . [S.l.: s.n.], 2015. Citado na página 20.

ALJALBOUT, E. et al. Clustering with deep learning: Taxonomy and new methods. *ArXiv*, abs/1801.07648, 2018. Citado na página 25.

ANDERSON, T. *The Theory and Practice of Online Learning*. 2nd. ed. Edmonton, AB, CAN: AU Press, 2009. ISBN 1897425082. Citado na página 13.

BALABANTARAY, R. C.; SARMA, C.; JHA, M. Document clustering using k-means and k-medoids. *CoRR*, abs/1502.07938, 2015. Disponível em: <http://dblp.uni-trier.de/db/journals/corr/corr1502.html#BalabantaraySJ15>. Citado na página 15.

BERRY, M. W. *Survey of Text Mining*. Berlin, Heidelberg: Springer-Verlag, 2003. ISBN 0387955631. Citado na página 14.

CASPI, A.; GORSKY, P.; CHAJUT, E. The influence of group size on nonmandatory asynchronous instructional discussion groups. *Internet and Higher Education*, Elsevier Ltd, v. 6, n. 3, p. 227–240, 2003. ISSN 1096-7516. Disponível em: <https://www.learntechlib.org/p/97095>. Citado na página 13.

CHOWDHURY, G. G. Natural language processing. *Annual Review of Information Science and Technology*, v. 37, n. 1, p. 51–89, 2003. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370103>. Citado na página 14.

Davies, D. L.; Bouldin, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, April 1979. ISSN 1939-3539. Citado na página 21.

DAY, W.; EDELSBRUNNER, H. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, v. 1, n. 1, p. 7–24, December 1984. Disponível em: <https://ideas.repec.org/a/spr/jclass/v1y1984i1p7-24.html>. Citado na página 21.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. [S.l.]: AAAI Press, 1996. (KDD'96), p. 226–231. Citado na página 21.

FERREIRA-MELLO, R. et al. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1332, 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1332>. Citado 4 vezes nas páginas 13, 14, 16 e 20.

Gerosa, M. A.; Fuks, H.; De Lucena, C. J. P. 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*. [S.l.: s.n.], 2001. p. 136–141. Citado na página 14.

GUO, H.; MA, J.; LI, Z. Active semi-supervised k-means clustering based on silhouette coefficient. In: XHAFA, F.; PATNAIK, S.; TAVANA, M. (Ed.). *Advances in Intelligent, Interactive Systems and Applications*. Cham: Springer International Publishing, 2019. p. 202–209. ISBN 978-3-030-02804-6. Citado na página 21.

HAN, J.; KAMBER, M.; PEI, J. *Data mining concepts and techniques, third edition*. Waltham, Mass.: Morgan Kaufmann Publishers, 2012. Disponível em: <http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1>. Citado na página 15.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, [Wiley, Royal Statistical Society], v. 28, n. 1, p. 100–108, 1979. ISSN 00359254, 14679876. Disponível em: <http://www.jstor.org/stable/2346830>. Citado na página 20.

KAUFMANN, L.; ROUSSEEUW, P. Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, p. 405–416, 01 1987. Citado na página 20.

KOVáCS, F.; LEGáNY, C.; BABOS, A. Cluster validity measurement techniques. *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, 02 2006. Citado na página 21.

LIM, L.-A. et al. What changes, and for whom? a study of the impact of learning analytics-based process feedback in a large course. *Learning and Instruction*, p. 101202, 2019. ISSN 0959-4752. Disponível em: <http://www.sciencedirect.com/science/article/pii/S095947521830450X>. Citado na página 16.

LOPEZ, M. et al. Classification via clustering for predicting final marks based on student participation in forums. *Proc. of 5th Int. Conf. on Educational Datamining*, p. 148–151, 01 2012. Citado na página 16.

MEANS, B. et al. Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. *Centre for Learning Technology*, Aug 2009. Citado na página 13.

MORILHAS, L. J. 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*, v. 1, n. 1, p. 66–88, Jan. 2009. Disponível em: <https://future.emnuvens.com.br/FSRJ/article/view/4>. Citado na página 13.

NASON, M. Learning together online: Research on asynchronous learning networks. *Education and Information Technologies*, v. 11, p. 191–192, 04 2006. Citado na página 13.

PARDO, A. et al. Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, v. 50, n. 1, p. 128–138, 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12592>. Citado na página 15.

PINHEIRO, A. et al. An analysis of the use of good feedback practices in online learning courses. In: . [S.l.: s.n.], 2019. Citado 3 vezes nas páginas 13, 16 e 25.

RAMOS, J. et al. A comparative study between clustering methods in educational data mining. *IEEE Latin America Transactions*, v. 14, p. 3755, 08 2016. Citado na página 16.

ROLIM, V.; MELLO, R. F.; LINS, R. D. Análise de discussões em fóruns educacionais usando mineração de texto e análise de grafos. *Sociedade Brasileira de Computação*, 2020. Citado na página 13.

Singh, V. K.; Tiwari, N.; Garg, S. Document clustering using k-means, heuristic k-means and fuzzy c-means. In: *2011 International Conference on Computational Intelligence and Communication Networks*. [S.l.: s.n.], 2011. p. 297–301. Citado na página 15.

WEVER, B. D. et al. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers Education*, v. 46, n. 1, p. 6 – 28, 2006. ISSN 0360-1315. Methodological Issues in Researching CSCL. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0360131505000552>. Citado na página 13.

WULF, J. et al. Massive open online courses. *Business Information Systems Engineering*, v. 6, p. 111–114, 02 2014. Citado 3 vezes nas páginas 13, 15 e 25.