UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ADRIANO EMANUEL GURGEL DE OLIVEIRA DOS SANTOS

MODELO PROBABILÍSTICO DE TÓPICOS E ESTATÍSTICA MULTIVARIADA APLICADOS À ANÁLISE TEXTUAL: um módulo de detecção de conversas fora do contexto para analisar conversas em grupo

Mossoró-RN

2020

ADRIANO EMANUEL GURGEL DE OLIVEIRA DOS SANTOS

MODELO PROBABILÍSTICO DE TÓPICOS E ESTATÍSTICA MULTIVARIADA APLICADOS À ANÁLISE TEXTUAL: um módulo de detecção de conversas fora do contexto para analisar conversas em grupo

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação - associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, para a obtenção do título de Mestre em Ciência da Computação.

Linha de Pesquisa: Tecnologias Aplicadas à Educação
Orientador: Carla Katarina Monteiro Marques, Profº Drª.
Coorientador: Francisca Aparecida Prado Pinto, Profº Drª.

Mossoró-RN

2020

Catalogação da Publicação na Fonte.
Universidade do Estado do Rio Grande do Norte.

G979m GURGEL DE OLIVEIRA DOS SANTOS, ADRIANO EMANUEL
MESTRE. / ADRIANO EMANUEL GURGEL DE OLIVEIRA DOS SANTOS. - Mossoró, 2020.
86p.

Orientador(a): Profa. Dra. CARLA KATARINA DE MONTEIRO MARQUES.
Dissertação (Mestrado em Programa de Pós-Graduação em Ciência da Computação). Universidade do Estado do Rio Grande do Norte.

1. Programa de Pós-Graduação em Ciência da Computação. I. DE MONTEIRO MARQUES, CARLA KATARINA. II. Universidade do Estado do Rio Grande do Norte. III. Título.

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC´s) foi desenvolvido pela Diretoria de Informatização (DINF), sob orientação dos bibliotecários do SIB-UERN, para ser adaptado às necessidades da comunidade acadêmica UERN.

ADRIANO EMANUEL GURGEL DE OLIVEIRA DOS SANTOS

MODELO PROBABILÍSTICO DE TÓPICOS E ESTATÍSTICA MULTIVARIADA APLICADOS A ANALISE TEXTUAL: UM MODULO DE DETECÇÃO DE CONVERSAS FORA DO CONTEXTO PARA ANALISAR CONVERSAS EM GRUPO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

APROVADA EM: 25 / 08 / 2020

Profa. Dra.Carla Katarina de Monteiro Marques
Orientador e Presidente

Dra. Francisca Aparecida Prado Pinto
Membro Externo - Secretaria da Educação do Estado de Ceará - Seduc

Prof. Dr. Henrique Jorge Amorim Holanda
Membro Externo - Universidade do Estado do Rio Grande do Norte - UERN

Prof. Dr. Paulo Ricardo Barboza Gomes
Membro Externo - Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE

Prof. Dr. Giovanni Cordeiro Barroso
Membro Externo - Universidade Federal do Ceará - UFC

Resumo

Fóruns de um Ambiente Virtual de Aprendizagem (AVA) é um sistema que proporciona a distribuição e o desenvolvimento de conteúdos diversos para cursos *online* e disciplinas semipresenciais para alunos em geral. Como um AVA é um ambiente virtual desenvolvido para ajudar professores e tutores no gerenciamento de conteúdos e materiais complementares para os seus alunos e na gestão completa de cursos online, é imprescindível que esse gerenciamento esteja relacionado ao que os alunos realmente discutem diante de um tema proposto à discussão. A dinâmica de discussão sobre um tema proposto cria uma enorme quantidade de dados na forma de texto, dificultando a tarefa de extrair conhecimento sobre as informações relacionadas. Visando minorar essa dificuldade, este trabalho faz uso da mineração de texto, por meio de duas técnicas tradicionais da estatística multivariada, LDA (do inglês, *Latent Dirichlet Allocation*) e PCA (do inglês, *Principal Components Analysis*), buscando verificar a eficiência e análise exploratória a fim de reportar a importância estatística dos termos analisados nos textos. Consegue-se, com este trabalho, não apenas realizar a redução da dimensão dos dados, como também a categorização, de modo automático, de documentos que estão na forma de dados textuais.

Palavras-chave: AVA, Fórum, Mineração de Texto, LDA, PCA.

Abstract

Forums of a Virtual Learning Environment (VLE) is a system that provides the distribution and development of diverse contents for online courses and semi presential subjects for students in general. As an AVA is a virtual environment developed to help teachers and tutors in the management of contents and complementary materials for their students and also in the complete management of online courses, it is essential that this management is also related to what the students really discuss when faced with a topic proposed for discussion. The dynamics of discussion on a theme proposed by the teacher creates in the forums an enormous amount of data available in the form of text and this makes it difficult, in a short time, the task of extracting knowledge about related information. To assist in this difficulty, this work makes use of text mining, through two traditional techniques of multivariate statistics, LDA (Latent Dirichlet Allocation) and PCA (Principal Components Analysis), seeking to verify the efficiency and exploratory analysis in order to report the importance statistical analysis of the terms analyzed in the texts. In this work, it is possible not only to reduce the size of the data, but also to automatically categorize documents that are in the form of textual data.

Keywords: VLE, Forums, Text Mining, LDA, PCA.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AVA	- Ambiente Virtual de Aprendizagem
QGP	- Questão Geral de Pesquisa
QC	- Questões Conceituais
QT	- Questos Tecnológicas
QP	- Questões Práticas
RSL	- Revisão Sistemática de Literatura
ESBE	- Engenharia de Software Baseada em Evidência
LDA	- Latent Dirichlet Allocation
PCA	- Principal Components Analysis
NB	- Naïve Bayes
KNN	- k-nearest neighbors
DT	- Decision Trees
SVM	- Support Vector Machine
RF	- Random Forest
PW-LDA	- Partitioned Word2Vec-LDA
ACC	- Accuracy
NMI	- Normalized Mutual Information
ARI	- Adjusted Rand Index
IFRN	- Instituto Federal do Rio Grande do Norte
BOW	- Bag of Words
PM	- Polícia Militar
RN	- Rio Grande do Norte

Sumário

1 Introdução

Neste capítulo é encontrado uma contextualização (Seção 1.1) acerca deste trabalho de dissertação, a proposta de estudo na Seção 1.2, os objetivos propostos na Seção 1.3, a organização do documento na Seção 1.4, as questões de pesquisa propostas na Seção 1.5, a metodologia na Seção 1.6 e motivação 1.7.

1.1 CONTEXTO

A proposta de grupos de estudo por meio dos fóruns de discussão vem se intensificando com uma mudança nas maneiras de se ensinar na sociedade contemporânea. Tal fato, deve-se à tecnologia que vem transcendendo a barreira de aulas presenciais e tornando o aprendizado uma maneira mais dinâmica. Nessa nova perspectiva, os grupos de estudo vêm ganhando força para que os alunos tirem dúvida, compartilhem seus conhecimentos e troquem ideias, facilitando o aprendizado. Segundo Alberti *et al.* (2014), a atividade de estudar em grupos dá orientação para construir atividades de estudo que sejam planejadas, monitoradas, conduzidas e avaliadas a fim de que o indivíduo que esteja envolvido nelas se desenvolva tanto em termos intelectuais como sociopsicológicos.

Com a observação de que os estudantes podem desviar o foco de determinados temas propostos em grupos de estudo, propõe-se, neste trabalho, descobrir estruturas temáticas em torno das quais os usuários (alunos) de uma ferramenta fórum de discussão de um AVA estão discorrendo. Isso possibilita testar técnicas da estatística que podem auxiliar docentes na hora de observar o contexto da conversa dos alunos. É muito difícil fazer a descoberta de padrões ou de conhecimento manualmente, por isso há a necessidade de alguma técnica que auxilie na extração de dados que seja relevante para esse contexto.

1.2 PROPOSTA DE ESTUDO

No AVA, os alunos são incitados a participarem de fóruns (grupos) de discussão o que gera uma grande quantidade de dados na forma textual. Esses dados, inicialmente, podem não representar nenhum conhecimento, mas, quando transformados, podem mostrar informações úteis subjacentes possibilitando a descoberta de padrões e ajudando a inferir algum comportamento. A proposta de analisar uma conversação de alunos de fórum de um AVA, com a utlização da mineração de dados, possibilita a compreensão da linguagem natural dos dados na forma textual e ajuda a lidar com a sua imprecisão e incerteza. Para conseguir transformar um texto em algo que o computador consiga entender, aplicar a mineração envolve várias áreas da informática, como mineração de dados (*data mining*),

aprendizado de máquina, recuperação de informação, estatística e linguagem computacional (SOUZA, 2016).

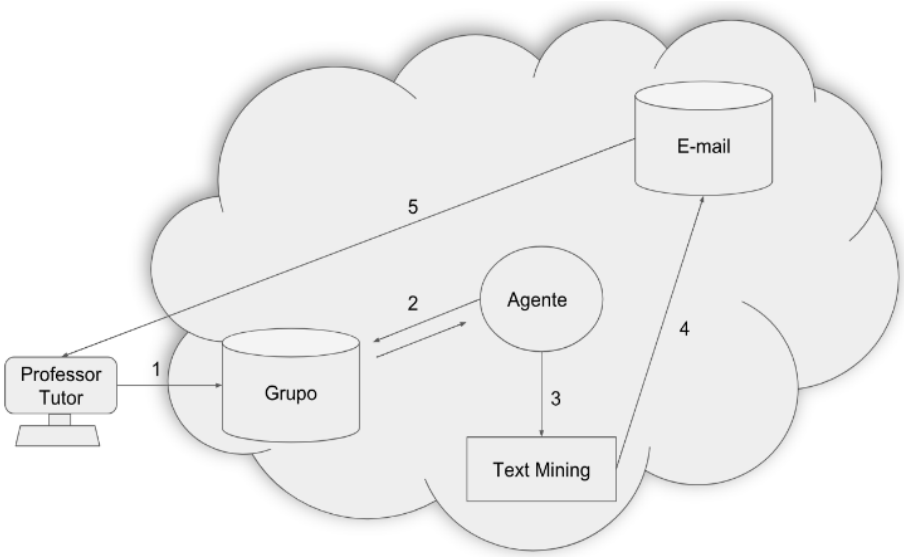
Assim, este trabalho visa à análise de uma conversação por meio de um artefato que use recursos estatísticos, respostas a questões de pesquisa, de uma revisão sistemática de literatura (RSL) e seguindo o *workflow* da Ciência de Dados, trazendo como consequência a comparação de técnicas da estatística, de modo que, categorize informações relevantes a serem reportadas ao docente/tutor, a fim de facilitar se há algum comportamento de desvio de padrão à discussão propsota.

Partindo do princípio de que a metaciência é uma ciência que desenvolve recursos e ideias para ela própia, dentro da proposta deste trabalho se tem a compreensão da resolução de um problema relevante sem a utilização de metaciência, ou seja, pretende-se com este trabalho usar os recursos que a estatística e a computação possibilitam ser acessados para gerar algo para um outro público que não seja a própria estatística ou computação. Dessa forma, pois, pretende-se gerar como benefício desse estudo um artefato que possui a tarefa de transformar dados fornecidos pelos alunos em informações úteis e que seja aplicado esse conhecimento como um recurso que ajude a área de tecnologias aplicadas à educação.

A Figura 1 apresenta uma visão geral dos elementos que contribuem para que exista uma mineração de texto com a utilização de agente (inteligente) em um grupo de estudo. Pode-se representar o fluxo do início em que o professor insere o tema de discussão até o próprio professor receber na sua caixa de entrada de e-mail uma mensagem avisando o andamento da conversa:

1. O professor preenche o formulário com o tema e indica a base de dados que deseja analisar;
2. Agente: monitora as conversas do fórum fazendo uma contagem de palavras relacionadas ao tema proposto pelo professor;
3. Módulo *text mining*: um algoritmo de mineração de texto recebe a permissão para começar a analisar os textos;
4. Armazena informação a ser enviada ao professor.
5. Envia informação ao professor.

Figura 1 – Elementos envolvidos na proposta inicial entre do professor e contexto dos alunos.



Fonte: Autoria Própria.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Este estudo tem como objetivo geral analisar se um texto produzido por um grupo de estudo, que discutem sobre um tema em um fórum de discussão de um AVA, está em conformidade com um determinado objetivo de aprendizagem.

1.3.2 Objetivos Específicos

1. Identificar fatores podem influenciar alunos a desviarem o contexto proposto;
2. Analisar dados na forma textual para que se consiga ter os termos mais relevantes dando controle ao docente sobre o andamentodo tema proposto;
3. Verificar a técnica mais eficiente ao contexto;
4. Analisar o efeito e a eficácia da mineração de texto aplicada na mediação entre docente e alunos.

1.4 ORGANIZAÇÃO DO DOCUMENTO

Este trabalho está organizado da seguinte maneira: o Capítulo 2 traz o referencial teórico, abordando assuntos inerentes à mineração de texto. Na sequência, o Capítulo 3 apresenta alguns trabalhos relacionados que objetivam o uso de uma técnica da mineração de

texto. O Capítulo 4 demonstra a aplicação proposta, sua estrutura e resultados alcaçandos por meio de respostas às questões de pesquisa. No Capítulo 5 são feitas as considerações finais do projeto. Logo em seguida vem o Capítulo 6, mostrando quais são os resultados esperados.

1.5 QUESTÕES DE PESQUISA

Para Alan *et al.* (2004), a *design science* é inerentemente um processo de resolução de problemas. O princípio fundamental desse tipo de pesquisa é que o conhecimento e a compreensão de um problema de *design* e sua solução são adquiridas na construção e aplicação de um artefato. Assim, este trabalho de dissertação possibilita questionamentos que, além de nortear para onde a pesquisa tem que ir, são respondidos ao se conseguir alcançar a direção dada pelas perguntas. Assim, o que se tem inicialmente são questões das quais ainda não se sabem as respostas, no entando se sabe quem pode responder.

Como a abordagem *design science* apresenta grande potencial para valorização da produção científica junto a sociedade em termos de aplicação, este trabalho adota o *design science* como paradigma de trabalho, ou seja, é uma metodologia adotada. Para Aken e Romme (2009), *design science* coloca a pesquisa de maneira mais atraente para outros pesquisadores das áreas das ciências que possuem forte vínculo prático no contexto da sociedade, direcionadas para resolução de problemas práticos.

A pesquisa desenvolvida neste trabalho de dissertação possibilita descobrir se os textos produzidos dentro um grupo de estudo, em um fórum de um AVA segue a proposta de discussão colocada por um professor. Por meio do estudo para essa descoberta, busca-se saber se há uma técnica de mineração de texto, mais eficiente possível, que possa ser aplicada a um módulo que faz essa mineração dentro de um fórum de discussão. As questões de pesquisa deste trabalho visa a perguntas como o porquê esse trabalho é relevante, para quem ele se destina, quando ele é necessário, como foi feito, passando por perguntas desde conceituais (QC), tecnológicas (QT) até práticas (QP). Assim, a questão geral de pesquisa (QGP) proposta é **como analisar se um texto produzido por um grupo de estudo na ferramenta fórum de um AVA está em conformidade com um determinado objetivo de aprendizagem**?

1. (QC) **Quais fatores podem influenciar alunos a desviarem o contexto proposto?**
    a) Quais dados devem ser utilizados e como extraí-los para desvendar as causas do desvio ao tema proposto?
2. (QT) **Como se deve analisar dados na forma textual para que se consiga ter os termos mais relevantes dando controle ao docente sobre o andamento**

do tema proposto?

a) Quais métodos são considerados relevantes para a análise de dados na forma textual?

b) Quais resultados da análise de dados textuais esses métodos reportam que podem ser considerados relevantes?

3. (QP) **Como analisar o efeito e a eficácia da mineração de texto aplicada na mediação entre docente e alunos?**

a) Quais os resultados da análise de desempenho, em termos de acurácia, ao aplicar as técnicas propostas no mesmo conjunto de dados?

b) Quais requisitos de aceitação e utilidade percebida do docente?

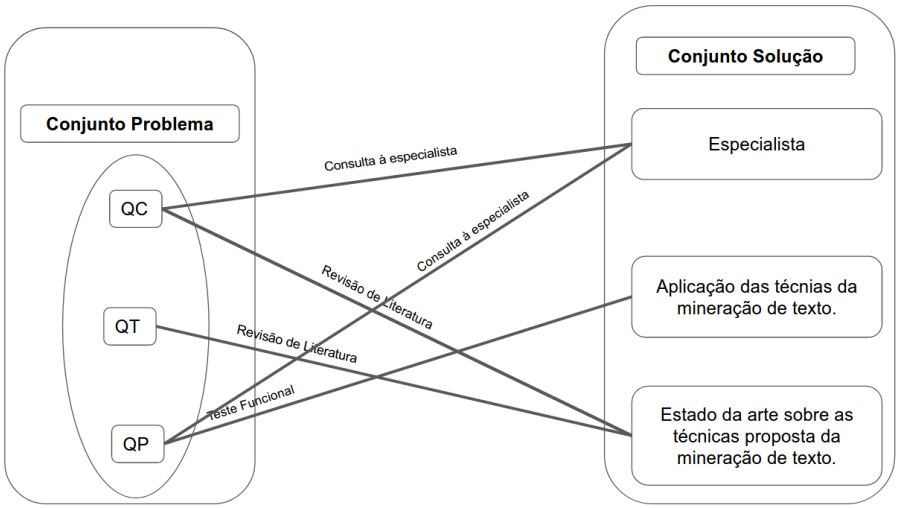
A estruturação descrita acima é o conjunto-problema de pesquisa, em *design science*, realizada através de uma Questão Geral de Pesquisa e suas subdivisões em Conceitual, Tecnológica e Prática. A elaboração do conjunto-solução ocorre pelas soluções apresentadas a essas questões. Para a aplicação *design science* é necessária a criação de um artefato inovador, que neste trabalho é o módulo de mineração de texto para um problema específico, isto é, o reconhecimento de padrões nos textos dos alunos em um fórum de um AVA. Dessa maneira, a questão tratada pode ser resolvida de maneira eficiente, utilizando-se um modelo estatístico devidamente pesquisado por meio de uma RSL e testado com resultados preliminares e finais.

1.6 METODOLOGIA

Dois métodos foram utilizados para obtenção de resultados para este trabalho. Inicialmente, foram desenvolvidas questões de pesquisa, por meio da *design science*, que possibilita a geração de um conjunto-problema a ser respondido mediante pesquisa e prática, conforme se verá ao longo desta dissertação. Outro método utilizado, porém na parte de desenvolvimento, é o *workflow* da Ciência de Dados, para aplicação dos métodos propostos da estatística como uma maneira de compará-los.

Conforme a Figura 2, a metodologia *design science* gera um conjunto-problema (QC, QT e QP) em que as soluções podem ser encontradas no conjunto-solução, ou seja, consegue-se observar que, para se responder a QC, pode-se utilizar a consulta a um especialista ou a revisão de literatura, em que esta gera o estado da arte, verificando ferramentas e aplicações diversas da utlização da mineração de dados. Neste trabalho é possível ver a aplicação desta metodologia na seção Revisão Sistemática de Literatura 3.1 e em Resultados 4.1.

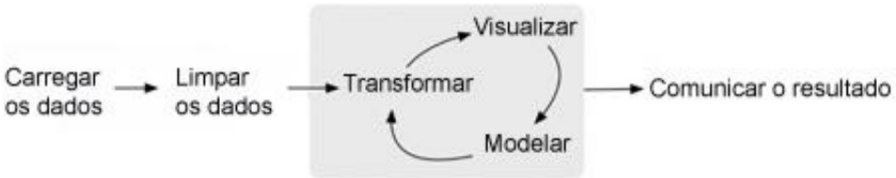
Figura 2 – Estrutura da pesquisa a partir do *design science*.



Fonte: Autoria Própria.

Na parte de desenvolvimento e testes das técnicas utilizadas para validar os recursos da estatística, foi utilizada a metodologia *workflow* da Ciência de Dados, desde a captura e a transformação dos dados até a comunicação dos resultados. Dessa forma, este trabalho visa encontrar um conjunto-solução, por meio da análise e interpretação de dados, que auxilie docentes na tarefa de investigar se um grupo de estudo está conversando sobre uma temática proposta. Este *workflow* é utilizado neste trabalho porque é a metologia seguida pela Ciência de Dados e esta é a área do conhecimento que se preocupa em estudar a parte de análise; possibilitando, pois, o estudo e na análise de dados, para extrair conhecimento e criar novas informações (OLIVEIRA; GUERRA; MCDONNELL, 2018). De acordo com Wickham e Grolemund (2016), um projeto típico em Ciência de Dados segue o *workflow* ilustrado na Figura 3.

Figura 3 – Ilustração de um workflow clássico utilizado em Ciência de Dados.



Fonte: Adaptado de Wickham e Grolemund (2016).

Este trabalho segue o *workflow* ilustrado na Figura 3, pois, na missão de auxiliar a tecnologia aplicada à educação, ver-se a necessidade de se utilizar recursos estatísticos. Assim, é por meio desse *workflow*, busca-se utilizar recursos estatísticos como uma maneira de comparar seus resultados e inferir sobre suas eficácias do ponto de vista da análise de desempenho. Para alcançar esse objetivo, faz-se necessária não apenas a obtenção dos dados, mas ainda a preparação deles para os modelos estatísticos. Como proposta

de se poder comparar esses modelos, na fase de **Carregar os dados**, trabalha-se com o mesmo conjunto de dados importados, em que inicialmente foi feita a fase de resultados preliminares 4.1.1, que ajudou na seleção da técnica que se apresenta mais eficiente ao problema proposto, na qual se trabalha com a base de dados do *Project Gutenberg* e, a posteriori, depois de validar as técnicas na fase anterior, foi realizada a obtenção de resultados mais consistentes para esta proposta com os resultados na base de dados do AVA 4.1.2. Dessa forma, para a utilização dos métodos da estatística, primeiro foi necessário escolher com qual base de dados se iria trabalhar para, depois, preparar-se os dados de entrada de forma igual para as técnicas de transformação propostas para se poder fazer um resultado com o qual se possa fazer um comparativo.

Conforme o passo **Limpar os dados** do $workflow$ da Ciência de Dados, é necessário limpar o texto antes de construir uma matriz de termos de documentos, que possui termos já pré-processados para análise textual.

A partir deste ponto em que os dados estão organizados, uma matriz de termos de documentos, gerada na fase de limpeza dos dados, realiza-se a fase de gerar conhecimento. Internamente os modelos estatísticos propostos realizam a fase **Transformar**, pois preparam os dados, criando novas variáveis e usando técnicas a fim de realizar os cálculos necessários para que alcancem os resultados esperados.

Depois de preparar os dados com as variáveis necessárias e aplicar a fase de transformação, entra-se na fase **Visualizar**, que possibilita a visualização da transformação e a geração do conhecimento através da interpretação do leitor sobre tabelas e gráficos (OLIVEIRA; GUERRA; MCDONNELL, 2018).

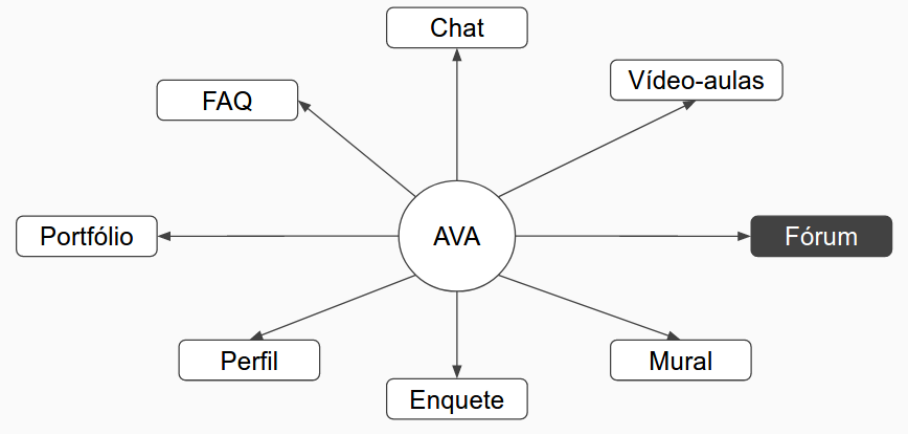
A fase de **Modelar** ajuda a responder as questões de pesquisa mostradas na introdução deste trabalho e será vista na seção Resultados 4.1, bem como a fase de **Comunicar** o resultado, em que há a discussão e difusão dos resultados.

1.7 MOTIVAÇÃO

A Educação à Distância (EAD) possui discussões em fóruns em que os usuários se comunicam por meio de dados não estruturados e formais na forma de texto; por conseguinte, esses dados podem ser analisados, pré-processados e gerando conhecimento que possibilite inferir sobre algum comportamento. Dessa forma, o auxílio à educação à distância, possibilita um professor/tutor a entender, mesmo sem observar diretamente as postagens dos alunos, sobre as ideias inseridas em um fórum de discussão.

Embora um AVA possua várias ferramentas, como pode ser vista na Figura 4, a motivação deste trabalho de dissertação está direcionada apenas a parte do fórum, pois é o local onde há possibilidade de inferir sobre qual tema os alunos estão discorrendo.

Figura 4 – Algumas ferramentas disponibilizadas por um AVA.



Fonte: Adaptado de Wickham e Grolemund (2016).

A motivação que fomenta este estudo, possibilita a análise de um grande universo de dados não estruturados na forma textual que pode existir em um fórum de discussão; logo, faz-se necessária a transformação de dados não útes inicialmente em dados úteis. Para isso, busca-se verificar se é possível a utilização recursos estatísticos e algorítmicos a fim de analisar se os alunos estão focados dentro do tema proposto.

2 Referencial Teórico

Quando se fala em Ciência de Dados, a área mais importante é o aprendizado de máquina (ou aprendizagem de máquina). Miller (2017) explica que esse é o processo que visa o programar um computador a fazer previsões realistas (ou melhorar as previsões), com base em algum fluxo ou fonte de dados. Sendo assim, obtém-se um processo contínuo e evolutivo, em que o torna semelhante a um ser vivo, por meio de sua relação com o ambiente. De modo mais detalhado, considera-se o aprendizado de máquina como um subcampo da ciência da computação, que evoluiu do estudo do reconhecimento de padrões e da teoria do aprendizado computacional, em inteligência computacional. Para Miller (2017), Arthur Samuel, em 1959, definiu o aprendizado de máquina como uma aŕea do conchecimento que possibilita os computadores a capacidade de aprender sem a necessidade de serem explicitamente programados. Nesse tipo de aprendizado, ao se buscar a extração de informações a partir de um *corpus*, facilita a execução de uma tarefa, que traria ônus se fosse executada manualmente. Isso implica em uma minuciosa análise dos textos do *corpus*. Dessa forma, o processo de descobrir tópicos está relacionado com a *clusterização* de termos que aparecem com frequência nas partes de texto do *corpus* que estão relacionadas. Blei, Ng e Jordan (2003) colocam que a computação tem esse problema computacional na modelagem de tópicos quando se busca descobrir a estrutura oculta que criou a coleção de texto observada.

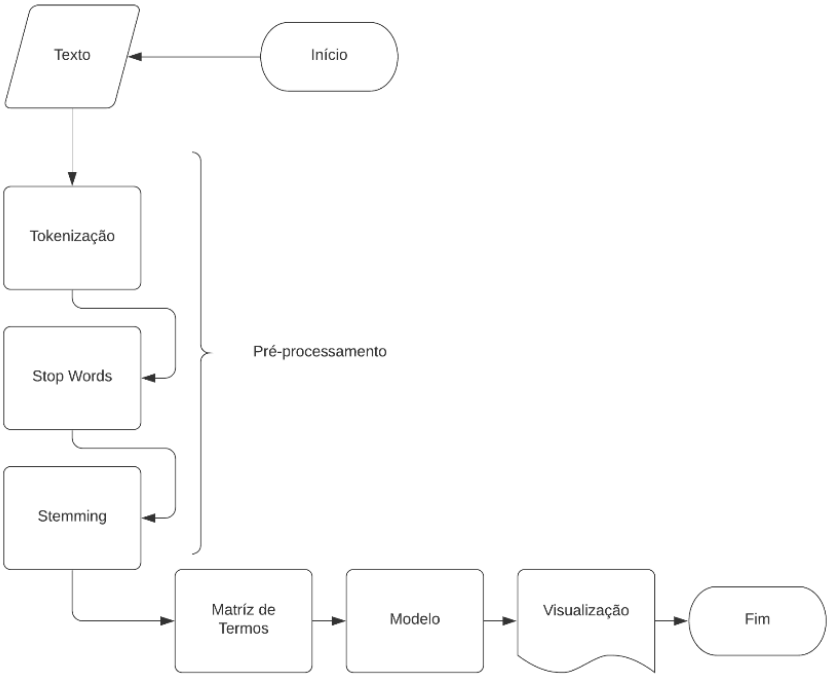
Para a fase de pré-processamento dos dados não estruturados, os dados textuais são carregados a partir da base de dados e concatenados conforme uma sequência temporal para se poder trabalhar em cima deles. Esses dados originais, carregados inicialmente, são chamados de *corpus*. Segundo Munzert *et al.* (2014), um *corpus* de texto em linguística simplesmente se refere a uma coleção estruturada de textos, ou seja, uma coleção estruturada de dados não estruturados na forma textual. Esse tipo de *corpus* utilizado aqui é, conforme Munzert *et al.* (2014), chamando de volátil, haja vista ser destruído quando se fecha a página. Existe também o *corpus* permanente, que é armazenado em um banco de dados.

Este capítulo apresenta os conceitos e terminologias básicas para a execução do desenvolvimento da aplicação direcionada à mineração de texto, dentro da proposta de estudo em um fórum de um AVA. Em primeiro lugar é apresentada a preparação dos dados e em seguida falamos sobre os modelos estatísticos propostos.

2.1 Pré-Processamento dos Dados

Para Sancheti, Shedge e Pulgam (2018), a redução de dimensão é uma etapa de pré-processamento na mineração de texto que transforma conjuntos de dados em um formato mais compacto e elimina redundância. Termos não úteis alteram a dimensão do conjunto de dados para um modelo de transformação na mineração de texto, fazendo com que a execução de um subconjunto de dimensões passe de um tempo de execução linear para um exponencial. Isso chama-se *curse of dimensionality* (maldição da dimensionalidade). Por isso, faz-se necessária a remoção de ruídos e redundância. Nesta fase de pré-processamento, primeiro há a separação dos termos em *tokens*, depois retira-se os *stop words*, aplica-se o *stemming* e cria-se uma matriz de termos de documento. Na Figura 5 é apresentada algumas fases utilizadas durante a pesquisa deste trabalho de dissertação, em que é possível ver a parte de pré-processamento, criação da matriz de termos e utilização do modelo:

Figura 5 – Fluxograma com as fases desde o pré-processamento até a apresentação dos resultados.



Fonte: Autoria Própria.

Cada passo da fese de pré-processamento pode ser entendida por meio da enumeração a seguir:

- *tokens*- neste trabalho, é a fase de tokenização em que o primeiro passo na fase de pré-processamento adotado é a colocação do texto no formato desejável a ser analisado dividindo as sentenças em *tokens*, ou seja, antes de se buscar algum padrão nas estruturas textuais, estas serão separadas em *tokens*, fazendo com que elas passem por um processo pelo qual são divididas em unidades de palavras. Essa etapa

é aconselhada ser usada antes de qualquer correspondência de padrão ser executada, haja vista que essa correspondência de padrões depende da existência de *tokens* (FELDMAN; SANGER, 2007).

- remoção de *stop words*- é a etapa da preparação dos dados que é adotada neste trabalho para se remover as palavras comuns da linguagem que geralmente não contribuem para a semântica dos documentos e não têm valor real agregado (FELDMAN; SANGER, 2007). No português brasileiro, há inúmeras palavras que são consideradas *stop words* como os verbos ser e estar, entre outros verbos, preposições, conjunções, artigos definidos e indefinidos, dentre outras.

- *stemming*- tem como objetivo principal reduzir a forma das palavras gramaticais a sua raiz (radical) (Ali; Khalid; Aslam, 2018). Para os autores Ali, Khalid e Aslam (2018), sistemas que processam dados textuais não estruturados são bem específicos de cada linguagem, com ampla aplicação a textos em inglês, constituindo uma etapa essencial de pré-processamento no processo de análise de texto e seu principal objetivo é reduzir as palavras gramaticais a sua forma.

2.2 Matriz de Termos de Documento

A principal razão para se reduzir essa dimensão de dados é a necessidade de transformar texto em números, gerando um BOW (*Bag of Words*) (MUNZERT *et al.*, 2014) em que cada palavra (termo) é um recurso ponderado num vetor. Esse peso pode ser gerado por um método TF (do inglês, *Term Frequency*, ou Frequência de Termos), IDF (do inglês, *Inverse Document Frequency* ou Frequência Invertida no Documento), TF-IDF ou valor binário. A enumeração a seguir, que pode ser vista na Figura 5 na fase **Matriz de Termos**, pode explicar melhor cada uma dessas técnicas:

- **TF** - conta o número de ocorrências (n) de um termo (t) em um documento (d). Como as palavras podem aparecer com mais frequência simplesmente porque os documentos são longos, pode-se normalizar a frequência dividindo-a pelo comprimento do documento, ou seja, tem-se a quantidade de ocorrências de termo em um documento dividida pela quantidade de termos exclusivos totais no documento (N). Assim, considerando n e N como variáveis quantitativas discretas e:
  - n sendo ocorrências do termo t em um documento d;
  - N número total de termos gerado a partir de d.

$$TF(d,t)\ retorna\ (n)/(N) \tag{2.1}$$

- **IDF** - mede a importância do termo t em todos os documentos (D), sendo D uma lista de documentos; obtém-se essa medida dividindo o número total de documentos (TD) pelo número de documentos contendo o termo (DCT), e depois, tomando o logaritmo desse quociente. Considerando TD e NCT como variáveis discretas quantitativas e:
  - TD sendo a quantidade total de documentos calculados a partir de D;
  - DCT número de documento que contém o termo t.

$$IDF(TD, DCT)\ retorna\ \log(TD/(DCT) \tag{2.2}$$

- **TF-IDF** - Esta técnica geralmente é utilizada em mineração de texto para converter entradas de texto em um vetor que contém o peso de cada termo em cada documento:

$$TF - ID(d, D, t)F\ =\ TF(d, t)\ *\ IDF(TD, DCT) \tag{2.3}$$

2.3 Modelos Estatísticos LDA e PCA

Para a utilização dos modelos LDA e PCA, como se pode observarvna fase **Modelo** na Figura 5, em um texto que precisa ser analisado para se descobrir algum comportamento que ali possa existir, é importante que a análise foque em como um computador pode entender. A aprendizagem, tanto para seres vivos como para uma máquina, é um processo contínuo e evolutivo. Os humanos, ao ler um texto, juntam as sentenças naturalmente em sua mente e formam os significados; já, para um computador, faz-se necessária a utilização de técnicas que inicialmente fazem um filtro (pré-processamento), preparando as palavras mais importantes, que podem trazer significado ao texto, para depois organizar os termos restantes uma matriz termo x documento (matriz de termo de documento). Após a preparação das palavras, pode-se ir para a fase de transformação da Ciência de Dados em que se propõe neste trabalho testes nos modelos LDA e PCA. A fim de que haja uma comparação entre esses dois modelos (ou técnicas), ambos irão trabalhar com a mesma fase de pré-processamento e mesma base de dados. Após a fase de pré-processamento e a preparação do BOW (*bag of words*), os dados ficam preparados para serem utilizados por cada modelo.

LDA e PCA são modelos estatísticos usados para ajudar a entender a variabilidade. Por variabilidade, entende-se como a distância que existe entre os dados, ou seja, quanto mais diferentes são, maior a variabilidade. Assim, sucessivas observações de textos podem não trazer os mesmos resultados, pois os textos representam fontes potenciais de variabilidade entre os termos (MONTGOMERY; RUNGER, 2012) devido a alterações seja entre conjuntos de dados, seja no mesmo conjunto, porém analisado por uma dimensão

diferente. Assim, a estatística desses métodos conseguem fornecer uma estrutura que descreve essa variabilidade e aprende quais fontes potenciais de variabilidade são mais importantes (MONTGOMERY; RUNGER, 2012). A diferença entre esses métodos em relação à variabilidade é que o PCA carrega a máxima variabilidade possível (Zare *et al.*, 2018).

Os modelos LDA e PCA são técnicas de apredizagem não supervisionadas (FALEIROS; LOPES *et al.*, 2016) e (JAMES *et al.*, 2013), em que os algoritmos têm que fazer por si só uma análise exploratória dos dados, buscando informações ocultas sobre os dados, e inferir a relevância dos termos. O LDA é um modelo probabilístico de tópico que simplica a exploração de grandes volumes de dados por meio da descoberta de tópcos, em que esses tópicos surgem a partir de uma análise dos textos originais (FALEIROS; LOPES *et al.*, 2016). O PCA se concentra como uma ferramenta para a exploração de dados não supersionados (JAMES *et al.*, 2013).

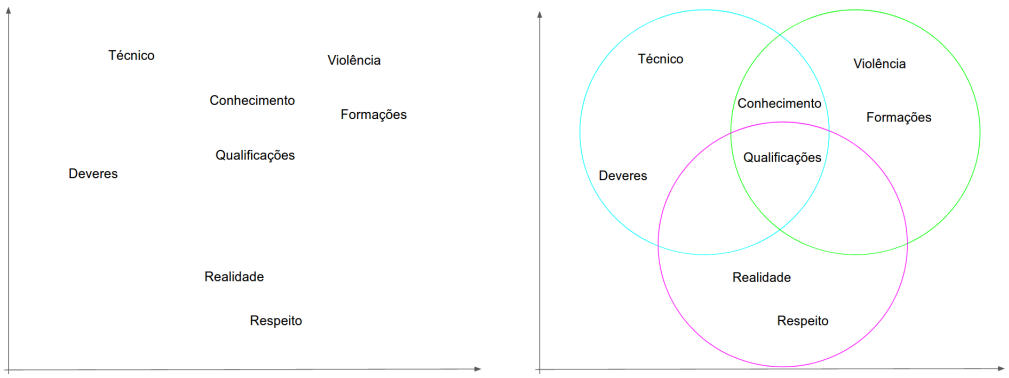
2.3.1 LDA

Como uma aplicação prática para este trabalho de dissetação, é imprescindível transformar uma grande quantidade de dados textuais advindo de um fórum em um AVA em informação útil. Como há o interesse em identificar os tópicos em textos do fórum, é possível considerar um método que não exija a especificação do número de textos que chegam no fluxo. Em um contexto não supervisionado, esses problemas caracterizam o agrupamento em fluxo de dados. Dessa forma, esse problema objetiva agrupar uma sequência X objetos em K grupos distintos, em que cada objeto deve ser atribuído a um dos K grupos na ordem que eles chegam no fluxo (BARTAL; CHARIKAR; INDYK, 1997). Como é possível ver na Figura 6, tem-se um exemplo de oito objetos distribuídos em três grupos. Assim, o problema de extração de tópicos em fluxo de dados textuais é considerado um caso especial do problema de agrupamento em fluxo de dados. Por meio de sua caractetística não supervisionada, o LDA faz busca por alguma semelhança matemática encontrada entre os termos para formar os agrupamentos (JAMES *et al.*, 2013) e, depois, categoriza esses agrupamentos em tópicos.

2.3.1.1 Tópicos

As palavras são utilizadas para formarem um texto. Obedecendo uma sequência lógica e semântica, elas representam um assunto e este pode ser considerado um tópico: subconjunto de palavras que acontecem com frequência nos textos e que assumem uma relação lógica e semântica, implicando um sentido dentro de um contexto. O plano cartesiano à direita da Figura 6 apresenta um exemplo de distribuição de palavras em três tópicos.

Figura 6 – Dados antes de clusterizar (esquerda) e clusterizados (direita).



Fonte: Autoria Própria

2.3.1.2 Modelos de Tópicos

A partir de um *corpus*, a modelagem de tópicos tem como objetivo a extração dos principais tópicos que representam os assuntos abordados nos textos do referido *corpus*. Dessa forma, esse tipo de modelagem em base de dados textuais é uma forma de mineração de texto. Por meio daquela, consegue-se identificar padrões latentes (escondidos) para encontrar a relação entre os textos e as palavras, ou seja, após a criação de um *corpus*, é possível a utilização de um algoritmo ou ferramenta que crie uma distribuição de tópicos. MIRIAM (2012) mostra que a modelagem de tópico é uma técnica para encontrar *clusters* de palavras, ou seja, tópicos em grandes corpos e texto. A Figura 6 apresenta uma modelagem de três tópicos.

A LDA é uma técnica da estatística pertencente aos modelos de tópicos em que a variável dependente é qualitativa (categórica) e essa variável é gerada a partir de variáveis independentes (RODRIGUES; COELHO; PAULO, 2007). No modelo LDA a variável dependente qualitativa é denominada tópico e depende de variáveis independentes que representam os termos do texto e que, além disso, estão dispostos aleatoriamente.

2.3.1.3 Modelagem Probabilística de Tópicos

Segundo Steyvers e Griffiths (2007), o modelo probabilístico de tópicos LDA é baseado na ideia de que documumentos são misturas de probabilidade de distribuição de tópicos e que tópicos são a probabilidade de distribuição de palavras. A Figura 6 apresenta três tópicos relacionados a um *corpus* (de texto). Cada um desses tópicos possuem uma probabilidade de pertencer a esse texto, assim como cada palavra associada a cada tópico possui uma probabilidade de pertencer a um tópico. As interseções da Figura 6 apresenta uma característica do modelo LDA: a clusterização difusa, em que pode haver palavras que têm probilidades de pertencer a mais de um tópico.

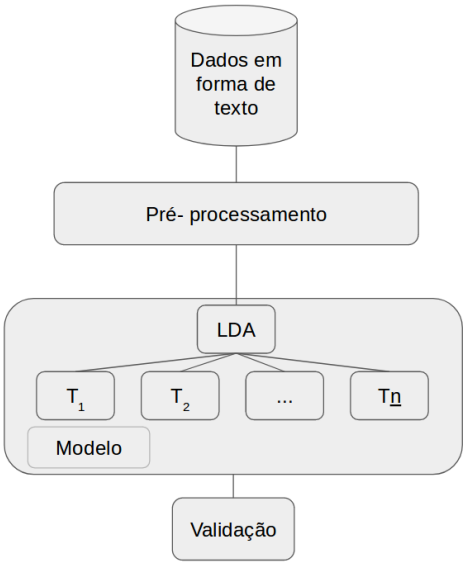
2.3.1.4 LDA

LDA é um método estatístico, mais especificamente é um modelo probabilístico, que pode ser usado para ajudar a entender a variabilidade. A variabilidade representa sucessivas observações de textos que podem não trazer resultados iguais, haja vista os textos representarem fontes potenciais de variabilidade (MONTGOMERY; RUNGER, 2012). Dessa forma, a estatística do LDA consegue fornecer uma estrutura que descreve essa variabilidade e aprende quais fontes potenciais dessa variabilidade são mais importantes e as organiza por similaridade.

Para o modelo LDA, depois de se agrupar os documentos para uma dimensão inferior por meio da categorização de tópicos, é possível aplicar outros algoritmos de aprendizado de máquina que se beneficiarão do menor número de dimensões. Naturalmente, o principal motivo pelo qual é utilizado esse modelo nesse trabalho de dissertação é que o LDA consegue, por meio dessa categorização, descobrir os temas ocultos e isso possibilitar descobrir temas relacionados a livros e a fóruns de discussão mediante um módulo de mineração de texto.

O modelo LDA (Figura 7) apresenta tópicos (T_1, T_2, ..., T_n) relacionados aos dados textuais pré-processados.

Figura 7 – Modelo pictórico para ilustrar as etapas da técnica LDA.



Fonte: Autoria Própria.

Para Steyvers e Griffiths (2007), um modelo generativo para documentos é baseado em regras simples de amostragem probabilística que descrevem como as palavras em documentos podem ser geradas com base em variáveis latentes. Esse modelo trabalha com o intuito de encontrar o melhor conjunto de variáveis latentes que explicam as palavras observadas em fóruns de discussão, ou seja, que explicam melhor os dados observados, assumindo que o modelo realmente gerou os dados categorizados em tópico.

Para entendermos a notação do LDA como um modelo probabístico, é preciso assumir que $P(z)$ significa a distribuição sobre os tópicos z em um documento específico e que $P(w|z)$ significa a distribuição de probabilidade sobre as palavras w, para um dado tópico z, conforme a função 2.4. Cada palavra w_i em um documento (referenciado pelo i-ésimo *token* de palavra) é gerada pela primeira amostragem de um tópico a partir da distribuição de tópico e, em seguida, a escolha de uma palavra é feita a partir da distribuição da palavra-tópico. Dessa forma, o LDA como modelo probabilístico trabalha com a seguinte distribuição de palavras dentro de uma estrutura de texto ananlisada:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \tag{2.4}$$

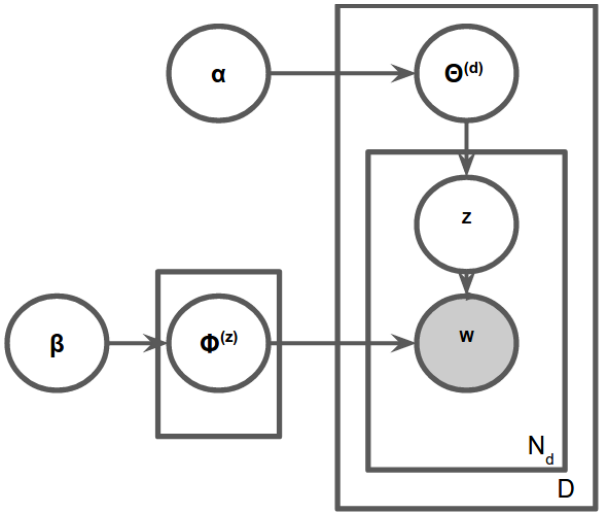
Em que T é o número de tópicos. Com o objetivo de simplificar a notação, vamos considerar que o parâmetro $\phi^{(j)} = P(w|z = j)$ represente a distribuição de palavras por tópico j e que o parêmetro $\theta^{(d)} = P(z)$ represente a distribuição tópicos por documento d. Como o modelo LDA pode trabalhar com vários documentos, assume-se que D é uma coleção de documentos e que cada documento d é constituído de N_d *tokens* de palavras. Assume-se também que N representa o número total de *tokens* de palavras, ou seja, $N = \sum N_d$. Os parâmetros ϕ e θ indicam quais palavras são importantes para um tópico específico e quais tópicos são importantes para um determinado documento.

Segundo Blei, Ng e Jordan (2003), LDA faz nenhuma suposição sobre como os pesos de mistura α são gerados, ajudando no teste da generalização do modelo. A densidade de probabilidade de uma distribuição Dirichlet T dimensional sobre a distribuição multinomial $p = (p_1, \ldots, p_T)$ é definida por:

$$Dir(\alpha_1, ..., \alpha_T) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_i^{\alpha_j - 1} \tag{2.5}$$

Os parâmetros dessa distribuição são especificados por $\alpha_1 ... \alpha_T$. Cada hiperparâmetro α_j pode ser interpretado como uma contagem de observação prévia para o número de vezes que o tópico j é amostrado em um documento, antes de ter observado quaisquer palavras reais daquele documento. Uma forma de ilustrar o LDA como um modelo generativo é por meio de uma rede Bayesiana conforme a notação da Figura 8.

Figura 8 – Representação gráfica de modelo de LDA. A caixa externa representa documentos, enquanto a caixa interna representa a escolha repetida de tópicos e palavras dentro de um documento.



Fonte: Adaptado de Blei, Ng e Jordan (2003)

Na Figura 8, as variáveis sombreadas são consideradas observáveis (termos de cada documento), enquanto aquelas que não estão sombreadas (distribuições de tópicos ou hiper-parâmetros) são consideradas latentes (não observáveis). As variáveis ϕ e θ, assim como z (que reprenseta a atribuição de *tokens* de palavras aos tópicos) são os três conjuntos de variáveis latentes que se deseja inferir. α e β são constantes nesse modelo. As setas indicam dependências condicionais entre variáveis, enquanto as placas (as caixas na figura) se referem a repetições de etapas de amostragem com a variável referente ao número de amostras. Por exemplo, as placas internas sobre z e w ilustram a amosragem da repetição de tópicos e palavras até que as palavras N_d tenham sido geradas para o documento d. A placa que envolve o $\theta^{(d)}$ ilustra a amostragem de uma distribuição sobre tópicos para cada documento d para um total de documentos em D. A placa em volta $\phi^{(z)}$ ilustra a amostragem repetida de distribuições de palavras para cada tópico z até que os tópicos T tenham sido gerado.

Na Figura 8, tem-se que cada um dos retângulos representa uma tarefa de repetição, em que a variável no seu interior rotula o número de vezes. Para Faleiros, Lopes *et al.* (2016), o modelo Bayesiano representado dessa forma apresenta-se em três níveis, desse modo a distribuição de tópicos para os documentos é feita pelo primeiro; a distribuição de tópicos para cada documento é realizada pelo segundo; a repetição da distribuição dos tópicos (internamente) às palavras de um documento é realizada pelo terceiro e esse último que faz a mistura de tópicos.

A partir deste ponto, em relação ao problema de inferência, podemos dar um passo da matemática e mapear "variáveis que conhecemos"versus "variáveis que não conhecemos":

- Parâmetros conhecidos
  - Documentos (d em D): Tem-se um número definido de documentos nos quais se identifica as estruturas de tópicos;
  - Palavras (w em W): Tem-se uma coleção de palavras e contagem de palavras para cada documento;
  - Vocabulário (W): a lista exclusiva de palavras em todos os documentos.
  - Hiperparâmetros:
    - α: Suposição prévia sobre a distribuição de tópicos de documentos. Fornece-se um valor α para inferência
      - Higher α: Assume-se que os documentos terão uma distribuição de tópicos semelhante e quase uniforme;
      - Lower α: Assumimos que as distribuições de tópicos do documento variam mais drasticamente.
    - β: Nossa suposição prévia sobre a distribuição de palavras de cada tópico.
      - Higher β: As distribuições de palavras em cada tópico estão mais próximas do uniforme, ou seja, cada palavra tem a mesma probabilidade de cada tópico;
      - Lower β: as distribuições de palavras variam mais de tópico para tópico.
- Parâmetros desconhecidos (latentes)
  - Número de tópicos (k): precisa-se especificar o número de tópicos que supomos que estejam presentes nos documentos. No entanto, não sabemos o número real de tópicos no *corpus*. Os métodos para estimar o número de tópicos em um *corpus* estão fora do escopo deste trabalho de dissertação.
  - Mistura de tópicos do documento (θ): precisamos determinar a distribuição do tópico em cada documento.
  - Distribuição de palavras de cada tópico (ϕ): Precisa-se conhecer a distribuição de palavras em cada tópico. Obviamente, algumas palavras vão ocorrer com muita frequência em um tópico, enquanto outras podem ter probabilidade zero de ocorrer em um tópico.
  - Atribuição de tópicos de palavra (z): Essa é realmente a principal coisa que se precisa deduzir. Para ficar claro, se há o conhecimento da atribuição de tópicos de cada palavra em cada documento, pode-se derivar a mistura de tópicos do documento θ e a distribuição de palavras ϕ de cada tópico..

Algoritmo 1: Pseudo-código processo gerador de um documento d

```
escolha θ Dirichlet (α)
inicio
    for cada uma das N palavras w_n do
        *
    end
fin
```

* escolha uma palavra w_n *de* $p(w_n|\theta, \beta)$

Os hiper-parâmetros α e β ajudam a determinar o comportamento dos tópicos. Quanto maior o valor de α, os documentos, muito provavelmente, estarão compreendidos com uma maior mistura de tópicos, e caso esse hiper-parâmetro tenda a ser menor a mistura compreenderá poucos tópicos. O hiper-parâmetro β tem um comportamento parecido, mas quando o valor de β cresce, os tópicos compreenderão uma maior probabilidae para ter mistura de várias palavras.

Um vetor bidimensional que representa a probabilidade do vocabulário é representado pela variável ϕ_k e ela é amostrada para cada um dos tópicos k. ϕ_k representa uma matriz n *x* K, em que as linhas indicam as palavras e as colunas, tópicos, conforme se por observar no exemplo da Tabela 1. Considerando a quantidade de tópcos K igual a 4 e a quantidade de palavras n igual a 3, é possível notar no exemplo que a soma das probabilidades de uma palavra de cada tópicos resulta em um.

Tabela 1 – Distribuição do vetor ϕ_k

Palavra x Tópico	k1	k2	k3	k4
w1	0.42	0,13	0,20	0,25
w2	00	0,65	0,30	0,05
w3	0,49	0,17	0,11	0,23

Associada a todos os documentos, temos a variável θ_j, em que θ é um vetor bidimensional m x K, onde as colunas representam os tópicos e as linhas, documento. θ_j representa um percentual de tópicos para um documento d_j de uma coleção, conforme Tabela 2.

Tabela 2 – Distribuição do vetor bidimensional θ_j

Palavra x Tópico	k1	k2	k3	k4
d1	0.48	0,13	0,30	0,09
d2	00	0,60	0,30	0,10
d3	0,39	0,27	0,11	0,23

Tem-se ainda que $z_{j,i}$ representa a distribuição de tópicos que está associado a

palavra $w_{j,i}$ em um documento d_j, em que $1 <= j <= m$ e $1 <= i <= n$. Ou seja, para as palavras, tem-se $z_{j,i}$ e $w_{j,i}$, são amostradas a cada palavra w_i para cada um dos documentos d_j. $z_{j,i}$ é a atribuição de um tópico k para a palavra w_i em documento d_j (FALEIROS; LOPES *et al.*, 2016).

Mais detalhes sobre a geração dos tópicos na modelagem LDA podem ser vistos em Blei, Ng e Jordan (2003).

2.3.2 PCA

A técnica PCA não possui variáveis dependentes, logo usa uma técnica da estatística multivariada que é a Análise Fatorial. Por meio da avaliação de um conjunto de variáveis, a Análise Fatorial, é uma técnica estatística que busca a identificação de dimensões de variabilidade comuns existentes em um conjunto de fenômenos, com o objetivo de desvendar estruturas não observáveis diretamente, em que cada uma dessas dimensões de variabilidade comum recebe o nome de fator (RODRIGUES; COELHO; PAULO, 2007).

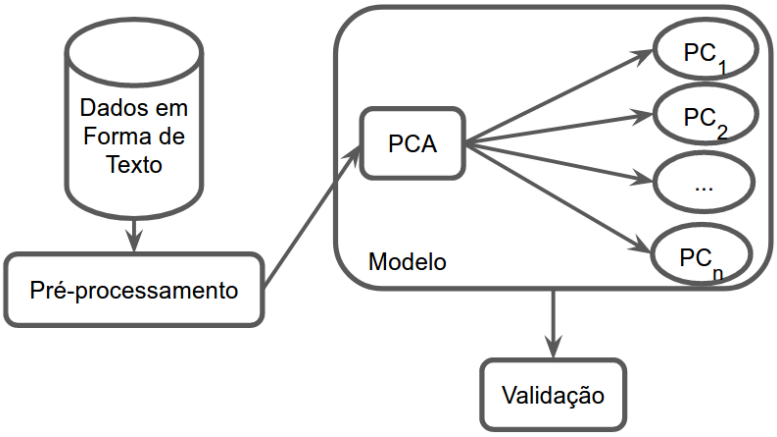
O método mais comum na Análise Fatorial é a *Principal Components Analysis*, por meio da qual se busca uma combinação linear entre as variáveis, de modo que essa combinação explique o máximo de variância. Seguidamente, retira-se a variância do passo anterior e procurar-se uma nova combinação linear entre as variáveis que explique a maior quantidade de variância restante, e assim por diante. Fatores ortogonais (não correlacionados entre si) são resultantes desse procedimento (RODRIGUES; COELHO; PAULO, 2007).

O nome Análise Multivariada está associado a um grande número de métodos e técnicas que utilizam, ao mesmo tempo, todas as variáveis na interpretação teórica do conjunto de dados obtidos (NETO, 2004). Como o interesse deste trabalho é verificar se um grupo de estudo está em conformidade com uma temática proposta, faz-se necessário verificar como as amostras se relacionam, bem como o quanto estas são semelhantes, que segundo as variáveis utilizadas, destaca-se como um dos objetivos de estudo o modelo PCA.

PCA é um dos métodos mais antigos e amplamente utilizados para redução de dimensionalidade em Ciência de Dados, cujo objetivo é encontrar os campos realmente importantes em bancos de dados com um grande número de variáveis, preservando a maior variabilidade possível (Zare *et al.*, 2018). Assim, os componentes são ordenados pela variância possível das variáveis e somente aquelas com maior variância são mantidos (MILLER, 2017).

O modelo PCA (Figura 9) retorna as coordenadas (PC_1, PC_2, ..., PC_n), que são as componentes principais e que carregam maior quantidade de informação no texto.

Figura 9 – Modelo pictórico para ilustrar as etapas da técnica PCA.



Fonte: Autoria Própria.

O modelo PCA transforma um conjunto de dados em combinações lineares tornando a interpretação mais agravável. Ele reduz o número de dados sem perda significativa de informação e, assim, facilita a interpretação dos dados. Para esse modelo, inicialmente, pega-se os dados originais e os transformam em dados transformados (combinações lineares), de modo que uma quantidade significativa de dados pode ser representada por poucas variáveis.

O formalistmo matemático do PCA coloca inicialmente os dados originais organizados em uma matriz, como se pode observar na Matriz 2.6, em que as amostras estão organizadas em linha e os parâmetros (dados originais), nos quais estão organizados essas amostras, estão organizados em coluna.

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix} \tag{2.6}$$

Na matriz supracitada se tem três observações e três amostras. Após organizar a matriz anterior, pode-se encontra a matriz de variância e co-variância, como se pode ver na Matriz 2.7.

$$\sum = \begin{bmatrix} Var_{(x1)} & Cov_{(x1x2)} & Cov_{(x1x3)} \\ Cov_{(x2x1)} & Var_{(x2)} & Cov_{(x2x3)} \\ Cov_{(x3x1)} & Cov_{(x3x2)} & Var_{(x3)} \end{bmatrix} \tag{2.7}$$

Depois de encontrada a Matriz 2.6, calula-se os autovalores e autovetores por meio da Equação 2.8.

$$det\left[\sum -\mu \right] = 0 \tag{2.8}$$

Ao aplicar a Equação 2.8, consegue-se encontrar os autovetores e autovalores. Os autovetores são os pesos de cada variável nas componentes principais, enquanto os autovalores são a infomração que cada componente principal carrega. Os autovalores estão representados pela Desigualdade 2.9.

$$\lambda_1 > \lambda_2 > \lambda_3 ... \lambda_p \tag{2.9}$$

Já os autovetores são representados pela Matriz 2.8.

$$\overline{a} = \begin{bmatrix} a_{i1} \\ . \\ . \\ . \\ a_{ip} \end{bmatrix} = 0 \tag{2.10}$$

Encontrados os autovetores e autovalores, consegue-se encontrar as componentes principais, conforme a Tabela 3. Para p parâmetros em que foram observadas as mostras, consegue-se achar as componentes principais, que são encontradas por meio da combinação linear entre os autovetores e autovalores.

Tabela 3 – Componentes Principais.

Componentes Principais	Funções
CP_1	$Y_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$
CP_2	$Y_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$
...	..
CP_p	$Y_p = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p$

3 Trabalhos Relacionados

Os trabalhos relacionados mostrados nesta pesquisa buscam evidenciar que há uma diferença ao se utilizar a técnica LDA para textos curtos que estão suscetíveis a ruídos e para textos formais (clássicos), mostrando qual a ideia proposta nesta dissertação ao se escolher os dados de um AVA. Os trabalhos relacionados envolvem desde técnicas de modelagem para textos curtos, até classificadores. Este Capítulo apresenta o uso das técnicas no que concerne à utilização dos métodos LDA e PCA, ou seja, apresenta técnicas que podem ou não sofrem uma alteração na sua estrutura interna devido a se trabalhar com textos organizados formalmente ou vir carregagos de gírias. Ao se trabalhar com dados de livros ou de um ambiente acadêmico, os dados conseguem ser bem rotulados e apresentados sem inconsistência devido a não se trabalhar com textos que possuem alguma má estrutura na sua confecção, como ruídos ou gírias conforme se pode observar no Capítulo *Os Artigos Relacionados* 3.2. Os trabalhos relacionados aqui apresentam claramente que quando não se tem uma estrutura formal na confecção do texto, é necessário realizar uma modificação do algoritmo para se alcançar resultados satisfatórios. Além disso, existem muitos trabalhos sobre mineração de texto que abordam técnicas diferentes que precisam de uma prévia seleção de termos e apresentam resultados com acurácia diferentes entre eles. Isso possibilita vislumbrar a literatura e analisar as diferentes soluções apresentadas.

Neste ponto, adianto que a RSL possibilitou não apenas a escolha do modelo estatístico LDA como a técnica ideal para ser aplicada ao módulo de mineração de texto proposto, como pode ser visto na seção Discussão 3.1.9, mas ainda possibilitou a escolha dos artigos relacionados.

3.1 Revisão Sistemática de Literatura

Atualmente, há uma quantidade de informação digital que propicia uma explosão na quantidade de dados, que são cada vez mais diversos, com maior complexidade, menos estruturados e que indicam uma necessidade maior de processamento mais rápido (PARK; NGUYEN; WON, 2015). Olhando para esse contexto, surge um paradigma chamado de *big data*, que é a existência de um grande volume de dados que se torna difícil coletar, armazenar e analisar, e possibilita também uma oportunidade para encontrar conhecimento útil (PARK; NGUYEN; WON, 2015) e, como a quantidade de dados na forma textual também vem crescendo de modo acelerado e proporcional aos outros tipos de dados, é cada vez mais importante desenvolver sistemas inteligentes para ajudar às pessoas a gerenciar e fazer uso de grandes quantidades de dados de texto ("*big text data*") (ZENG; CHEUNG; LIU, 2012). Dentro do universo do *big data*, há alguns conceitos como mineração

de dados, que é a descoberta de conhecimento dentro de um determinado conjunto de dados; e mineração de texto, que é uma parte do *data mining*, que se propõe a descobrir conhecimento em dados na forma de texto (OLIVEIRA; GUERRA; MCDONNELL, 2018).

Sabendo-se que uma RSL é uma técnica relevante na Engenharia de Software Baseada em Evidência (ESBE) idealizada por Kitchenham (2004), por meio dela pode-se identificar, avaliar e interpretar toda a pesquisa disponível relevante para uma questão de pesquisa em particular, ou área temática, ou fenômeno de interesse de modo imparcial e repetido (KITCHENHAM, 2004). Diferente da revisão bibliográfica tradicional, a revisão sistemática aqui utilizada segue um protocolo, ou seja, um passo a passo que é rigosamente obedecido, a fim de se alcançar na literatura artigos que tenham qualidade sobre a temática aqui em discussão. Assim, dentro de um grande universo de dados na forma textual não estruturados, em que existe dificuldade para se identificar algum comportamento nos textos, busca-se investigar trabalhos sobre a mineração de texto que utilizam o modelo probabilístico de tópicos por meio da aplicação da técnica LDA e da estatística multivariada por intermédio da técnica PCA.

Como a mineração de texto é a descoberta de conhecimento interessante em documentos de texto, ela se apresenta como uma área de pesquisa que tenta resolver problemas em informações textuais sobrecarregadas fazendo uso de técnicas de *data mining*, aprendizado de máquina, processamento de linguagem natural (PLN), recuperação de informações (RI) e gerenciamento de conhecimento (FELDMAN; SANGER, 2007). Assim, por meio da mineração de texto, nesse processo intenso de descoberta de conhecimento em textos não estruturados, esta RSL faz busca por estudos primários das técnicas LDA (do inglês, *Latent Dirichlet Allocation*) e PCA (do inglês, *Principal Components Analysis*) da estatística.

Evidenciando a existência de *big data*, mineração de dados e mineração de texto, mostra-se a importância de se buscar trabalhos na literatura que ajudem na pesquisa sobre as técnicas de mineração de texto que se mostram como uma solução na hora de fazer a descoberta de conhecimento. Desse modo, esta RSL investiga trabalhos existentes de maneira organizada por processo sistemático de pesquisa, por meio de uma estratégia de busca pré-definida no protocolo de revisão. Essa estratégia deve permitir que a integralidade desta pesquisa possua qualidade. Entre as contribuições desta RSL, além de buscar trazer valor científico mostrando resultados na aplicação das técnicas LDA e PCA, mostram-se os seus benefícios e limitações. Propõe-se mostrar as técnicas utilizadas como solução nos artigos selecionados. Além de propor também um enquadramento para posicionar adequadamente novas atividades de pesquisa. Por meio dos resultados, benefícios, limitações e soluções, esta RSL mostra se é possível a utilização de modelos estatísticos complexos, por meio das técnicas LDA e PCA para a análise textual, bem como comparar os resultados da análise de desempenho, em termos de acurácia, ao aplicar as duas técnicas no mesmo

conjunto de dados. Dessa forma, sabendo-se que o propósito principal deste trabalho de dissertação é a descorberta de qual técnica é a melhor para se aplicada em um módulo de mineração de texto, a intenção da RSL aqui é vislumbrar sobre os métodos LDA e PCA por meio execução de um protocolo de revisão que permite a execução de uma *string* de busca para seleção dos artigos que envolvem palavras-chave e sinônimos.

Este estudo secundário, no caso a RSL aqui utilizada, não é considerada uma revisão de literatura convencional, pois, para ser iniciado, é necessária a definição de um protocolo de revisão que especifica as questões de pesquisa sobre os modelos LDA e PCA. Outrossim, faz-se necessária, também, para este estudo, uma estratégia de busca bem definida que visa à detecção de grande parte da literatura relevante quanto possível sobre os modelos estatísticos propostos em bases de dados reconhecidas da computação, pois é onde se consegue resultados bem consistentes sobre cada técnica. Toda estratégia de busca aqui realizada é documentada para que os leitores possam acessar com rigor e integridade todo processo de desenvolvimento da pesquisa. Além disso, há a necessidade de se ter critérios explícitos de inclusão, exclusão e qualidade para avaliar um potencial estudo primário encontrado.

3.1.1 Protocolo da Revisão

Esta RSL é utilizada para se realizar buscas por estudos individuais que tragam contribuição importante à pesquisa (KITCHENHAM, 2004). Para Kitchenham (2004), esses estudos individuais que contribuem para uma revisão sistemática são chamados estudos primários.

Esta revisão sistemática segue o processo recomendado por Biolchini *et al.* (2007). Biolchini *et al.* (2007) usa algumas diretrizes definidas por Kitchenham (2004). Dois pesquisadores participaram do processo de revisão: um revisor principal e um pesquisador sênior responsável por validar a revisão. Dessa forma, este protocolo de revisão especifica os métodos que serão utilizados para realizar esta revisão sistemática. Para conduzir a RSL desta pesquisa, primeiro é realizada a definição dos objetivos e escopo. A segunda parte da execução é realizada com a definição das questões de pesquisa. O terceiro passo a ser executado é a definição das estratégias de busca. A quarta etapa da execução deste protocolo é a definição da seleção dos estudos. A quinta etapa é a definição dos procedimentos para seleção dos estudos e, por fim, é realizada a extração de dados.

3.1.2 Objetivos e Escopo

É importante observar que existem os objetivos (1.3) delineados para este trabalho e os objetivos que se pretende com a RSL. Assim, as motivações para esta revisão sistemática surgem a partir do interesse de investigar a utilização dos métodos LDA e PCA para

realizar reconhecimento de padrão em texto a fim de se conseguir alcançar o objetivo geral que é analisar se os alunos de um fórum de um AVA estão dentro do tema proposto por um docente para a discussão. Consegue-se, assim, auxiliar pesquisadores que tenham interesse pela área de processamento de linguagem natural. A fim de se buscar o método mais eficiente de reconhecimento de padrão em texto, são planejados os seguintes objetivos:

- **Objetivo 1**: Identificar como os métodos LDA e PCA podem ser utilizados para a extração de comportamento semântico em textos;
- **Objetivo 2**: Identificar estudos experimentais que foram conduzidos para validar esses métodos;
- **Objetivo 3**: Comparar os métodos LDA e PCA em relação à eficiência.

A fim de delimitar o escopo, pode-se relacionar os seguintes itens a esta pesquisa:

- **Intervenção**: quando artigos que contenham as palavras-chave ou sinônimos da *string* de busca são encontrados, o revisor principal desta RSL verifica a sua relevância para esta pesquisa.
- Controle:
  - PARK, Kyounghyun; NGUYEN, Minh Chau; WON, Heesun. Web-based collaborative big data analytics on big data as a service platform. 2015 17th International Conference On Advanced Communication Technology (icact), [s.l.], p.1-1, jul. 2015. IEEE;
  - FENG, Shuchao; SHANG, Wenqian; WANG, Yuqi. A k-Highest Expert Text Classification Algorithm Based on Choquet Integral. 2015 3rd International Conference On Applied Computing And Information Technology/2nd International Conference On Computational Science And Intelligence, [s.l.], p.11-11, jul. 2015. IEEE; http://dx.doi.org/10.1109/acit-csi.2015.95.
  - LI, Yuefeng et al. Relevance Feature Discovery for Text Mining. Ieee Transactions On Knowledge And Data Engineering, [s.l.], v. 27, n. 6, p.1656-1669, 1 jun. 2015. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tkde.2014.2373357;
  - ZHONG, Ning; LI, Yuefeng; WU, Sheng-tang. Effective Pattern Discovery for Text Mining. Ieee Transactions On Knowledge And Data Engineering, [s.l.], v. 24, n. 1, p.30-44, jan. 2012. Institute of Electrical and Electronics Engineers (IEEE).
- **População**: pesquisadores da área de estatística que têm interesse na modelagem de tópicos e na estatística multivariada aplicada à mineração de texto;

- **Resultados**: verificar a eficiência entre as técnicas LDA e PCA;
- **Aplicação**: pesquisadores.

3.1.3 Questões de Pesquisa

É importante que seja notado que há as questões de pesquia pensadas para o trabalho de dissertação e as questões de pesquisa pensadas para a RSL. Para atender aos objetivos propostos, desenvolveu-se questões de pesquisa que visa ao que o estado da arte possui sobre os métodos LDA e PCA no reconhecimento de padrão em texto.

- **Questão primária (QP1)**: Como os métodos LDA e PCA podem ser utilizados para a extração de comportamento semanticamente em textos?
- **Questão secundária (QS1)**: Qual estudo experimental conduzido para validar os métodos LDA e PCA?
- **Questão secundária (QS2)**: Em termos de eficiência, qual método (LDA ou PCA) apresenta melhores resultados?

3.1.4 Estratrégia de Busca

A estratégia de busca para seleção dos estudos primários, consiste na definição de critérios de escolha das fontes de busca, estratégia de pesquisa, palavras chaves, *string* de busca, definição dos tipos de estudos primários de interesse e idioma de escrita dos estudos, conforme apresenta-se a seguir:

- **Critério de Seleção das Fontes**: bases de textos indexadas mais conhecidas na área de ciência da computação;
- **Bases de Dados**: IEEE Xplore, ACM Digital Library, Science Direct, como se pode ver na Tabela 4;
- **Estratégia de Pesquisa**: artigos foram pesquisados em bibliotecas digitais relevantes na área de computação, sendo priorizados os anos entre 2009 e 2020. A Tabela 4 mostra as revistas e conferências selecionadas. As fontes foram selecionadas por serem consideradas importantes bases de divulgação de pesquisas internacionais, e por incluírem tanto trabalhos práticos quanto teóricos. Essas bibliotecas permitem consultas *on-line* por meio de mecanismos de busca, nos quais é possível utilizar expressões lógicas para definir a *string* de busca. Selecionou-se que somente os trabalhos no idioma Inglês seriam considerados nesta Revisão Sistemática.

Tabela 4 – Bases de dados utilizadas para a pesquisa.

Fonte	**Acrônimo**
Institute of Electrical and Electronics Engineers	IEEE Xplore Digital Library
Association for Computing Machiner	ACM Digital Library
Science Direct	-

Fonte: Autoria Própria.

A partir da combinação das palavras-chave e seus sinônimos, a *string* de busca desta RSL foi desenvolvida. Pode-se ver essas palavras na Tabela 5 em inglês e Tabela 6 em português.

Tabela 5 – Palavras-chave e sinônimos em inglês.

Palavras-chave	**Sinônimos**
LDA	Latent Dirichlet Allocation
PCA	Principal Components Analysis
Text Mining	Text Analysis

Fonte: Autoria Própria.

Tabela 6 – Palavras-chave e sinônimos em português.

Palavras-chave	**Sinônimos**
ADL	Alocação de Dirichlet Latente
ACP	Análise de Componentes Principais
Mineração de Texto	Análise de Texto

Fonte: Autoria Própria.

Em cada fonte de pesquisa da Tabela 4 foi utilizada a opção de busca avançada, os filtros disponibilizados para a selação dos tipos de artigos e o período entre datas de publicação. Como cada base de dados de pesquisa possui suas particularidades de busca avançada, a *string* de busca em cada uma delas sofreu adaptações.

3.1.5 Estratégia de Seleção dos Estudos

Esta subseção mostra não apenas a avaliaçao de qualidade como é feita, como ainda os seguintes critérios de inclusão e exclusão definidos para este estudo:

- **Critérios de Inclusão (CI1)**: apresenta identificação de comportamento semanticamente em textos por meio dos métodos LDA ou PCA;

- **Critérios de Inclusão (CI2)**: apresenta como o estudo experimental foi conduzido para validar os métodos LDA ou PCA;
- **Critérios de Inclusão (CI3)**: apresenta um estudo comparativo com as técnicas LDA ou PCA;
- **Critério de Exclusão (CE1)**: não apresenta identificação de comportamento por meio das técnicas LDA ou PCA;
- **Critério de Exclusão (CE2)**: não apresenta estudo experimental conduzido para para as técnicas estudadas neste trabalho;
- **Critério de Exclusão (CE3)**: trabalhos científicos repetidos.

Como parte dos critérios de inclusão/exclusão, esta seção também determina características pelas quais os trabalhos podem ou não ser selecionados para esta RSL. Com base na Diretriz CRD, Kitchenham (2004) sugere o uso de uma avaliação como um modelo de estudo para garantir o nível mínimo de qualidade, de modo a buscar-se o propósito de selecionar os melhores artigos para esta revisão, em que foram aplicados, também, critérios com o objetivo de avaliar a qualidade dos artigos encontrados. Esses critérios são:

- **Critério de Qualidade 1 (CQ1)**: Artigos publicados no período de 2009 a 2020;
- **Critério de Qualidade 2 (CQ2)**: Trabalhos que utilizam o idioma inglês;
- **Critério de Qualidade 3 (CQ3)**: Trabalhos que especificam a forma de coleta dos dados;
- **Critério de Qualidade 4 (CQ4)**: Trabalhos que declaram claramente uma técnica de mineração de texto investigada;
- **Critério de Qualidade 6 (CQ6)**: Livros da área de mineração de dados ou mineração de texto.

3.1.6 Procedimento para Seleção dos Estudos

Mediante uma busca por trabalhos realizada manualmente nas fontes de pesquisa apresentadas na Tabela 4, foi executado o processo de seleção dos estudos. Seguindo as regras de cada base de dados, as *strings* foram construídas, por meio da combinação das palavras-chave, seus sinônimos e pela utilização do campo de busca avançada. Após a execução da *string* de busca nas bases de dados, guardou-se os artigos retornados. Para fazer uma melhor seleção dos artigos, na segunda etapa de seleção, foi realizada a exclusão de artigos repetidos, leitura do título e resumo, aplicando-se, concomitantemente, os

critérios de inclusão e exclusão. Em seguida, com os artigos restantes, que são considerados relevantes, foi realizada a leitura completa, verificando-se ainda se os artigos obedecem a algum dos critérios de inclusão ou exclusão a fim de serem aceitos para esta pesquisa.

3.1.7 Extração de Dados

Conforme Kitchenham (2004), o formulário de extração de dados (Figura 10) desta RSL é projetado para coletar todas as informações que abordam as questões de revisão. Dessa forma, a projeção do formulário de extração de dados deste estudo visa mostrar a pesquisadores que vislumbram a mineração de texto sobre as informações que se podem obter com os estudos primários desta pesquisa.

Figura 10 – Itens e descrição do formulário de extração de dados.

FORMULÁRIO DE EXTRAÇÃO DE DADOS			
Título:			
ID de Estudo	**Autores**	**Fonte**	**Nº de Páginas**
Data da Publicação	**Idioma**		**Técnica Utilizada** ☐ LDA ☐ PCA
Objetivo			
Questão de Pesquisa Respondida ☐ Questão primária (**OP1**) ☐ Questão secundária (**QS2**) ☐ Questão secundária (**QS3**)			
Validação			
Observações			

Fonte: Autoria Própria.

3.1.8 Condução da Revisão

Por meio de uma *string* de busca construída com palavras-chave e sinônimos associados a área da mineração de texto juntamente com as áreas da modelagem de tópicos e da estatística multivariada com os métodos LDA e PCA respectivamente, foi executada inicialmente uma pesquisa na base de dados *IEEE Xplore* e depois essa *string* de busca foi adaptada para outras bases de dados reconhecidas na computação. Na base de dados *IEEE Xplore*, foram realizados vários testes para se conseguir desenvolver uma *string* de

busca que melhor se encaixe com esta pesquisa, como se pode observar abaixo:

- IEEE Xplore
  - String de busca ((("PCA"*OR* "LDA") *AND* ("text"*AND* "mining"*AND* "document*")));
    - 13 registros;
    - aplicados a *Journals* e *Magazines*;
    - 2010-2018, no IEEE Xplore.

- ACM Digital Library
  - String de busca ((("PCA"*OR* "LDA") *AND* ("text"*AND* "mining"*AND* "document*")));
    - 35 registros;
    - aplicados a *Workshop* e *Conferences*;
    - resultados com data a partir de 2009.

- Science Direct
  - String de busca (("*Latent DirichletAllocation*"*OR* "*Principal Components Analysis*"*OR* "PCA") *AND* ("*text mining*"*OR* "*text analysis*"));
    - 13 registros
    - aplicados a *Pattern Recognition* e *Journal of Multivariate Analysis*;
    - resultados a partir de 2009.

A Tabela 7 mostra os detalhes da pesquisa com a seleção de estudos por base de dados. As Tabelas 8, 9 e 10 mostram mais detalhadamente como foi realizada a selação para os registros de cada base de dados.

Tabela 7 – Detalhes da pesquisa e seleção de estudos por base de dados.

Bases de Dados	**Resultados da Pesquisa**
IEEE Xplore	13
ACM Digital Library	35
Science Direct	13
Encontrados	61
Incluídos	32
Excluídos	29

Fonte: Autoria Própria.

Com o resultado das pesquisas realizadas por meio dos mecanismos de busca avançada de cada fonte de pesquisa, obteve-se um total de 61 artigos, conforme Tabela 7 e Figura 11. Após a leitura dos títulos e resumos, ficaram ao todo 32 trabalhos restantes, ver Figura 12. Esses trabalhos passaram pela fase de pré-seleção e foram lidos integralmente. Após a leitura completa, 12 artigos ficaram no final e foram os escolhidos, através de um processo sistemático, para a extração de dados. Dessa forma, passando pelos critérios sistemáticos do protocolo da revisão sistemática, 12 trabalhos se mostraram relevantes e foram selecionados para a extração de dados, como se pode ver em detalhes nas tabelas Tabela 8, Tabela 9 e Tabela 10, e como mostrado de forma mais geral na Tabela 11, em que estão separados por base de dados em que foram encontrados.

Tabela 8 – Detalhes da pesquisa e seleção de estudos na base de dados IEEE Xplore.

Situação	**Quantidade**
Encontrados	13
Incluídos	4
Excluídos	9

Fonte: Autoria Própria.

Tabela 9 – Detalhes da pesquisa e seleção de estudos na base de dados ADM.

Situação	**Quantidade**
Encontrados	35
Incluídos	7
Excluídos	28

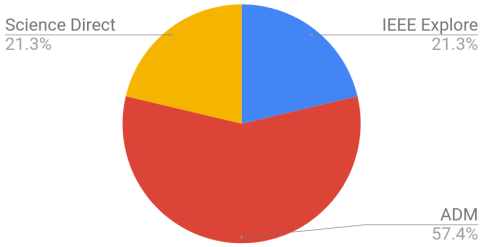
Fonte: Autoria Própria.

Tabela 10 – Detalhes da pesquisa e seleção de estudos na base de dados *Science Direct*.

Situação	**Quantidade**
Encontrados	13
Incluídos	1
Excluídos	12

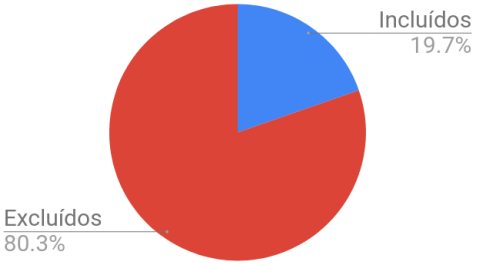
Fonte: Autoria Própria.

Figura 11 – Seleção de estudos por base de dados.



Fonte: Autoria Própria.

Figura 12 – Seleção de estudos por inclusão/exclusão.



Fonte: Autoria Própria.

Tabela 11 – Detalhes da pesquisa e seleção de estudos por base de dados.

Bases de Dados	**Resultados dos Critérios de Inclusão e Exclusão**
IEEE Xplore	4
ACM Digital Library	7
Science Direct	1
Total	12

Fonte: Autoria Própria.

A Tabela 12 mostra os artigos considerados relevantes para esta pesquisa por meio do título, autores, ano e a técnica pesquisada que o artigo aborda.

Tabela 12 – Artigos encontrados e suas bases de dados.

Ordem	**Título**	**Autor**	**Ano**	**Técnica Abordada**	**Base de Dados**
1	Weakly Supervised Joint Sentiment-Topic Detection from Text	C. Lin; Y. He; R. Everson e S. Ruger	2012	LDA	IEEE Xplore
2	Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents	M. ALHAWARAT1 M. HEGAZI	2018	LDA	IEEE Xplore
3	Learning Topic Models by Belief Propagation	Jia Zeng, Member; William K. Cheung; Senior Member and Jiming Liu	2012	LDA	IEEE Xplore
4	Extracting parallel fragments from comparable documents using a generative model	Somayeh Bakhshaei; Reza Safabakhsh e Shahram Khadivi	2018	LDA	SCIENCE DIRECT
5	Effective text classification using multi-level fuzzy neural network	Shima Zobeidi ; Marjan Naderan ; Seyed Enayatollah Alavi	2017	PCA	IEEE Xplore

6	A Non-Parametric Topic Model for Short Texts Incorporating Word Coherence Knowledge	Yuhao Zhang; Wenji Mao; Daniel Zeng;	2016	LDA	ACM DIGITAL LIBRARY
7	Probabilistic Topic Models for Text Data Retrieval and Analysis	ChengXiang Zhai	2017	LDA	ACM DIGITAL LIBRARY
8	Language usage on Twitter predicts crime rates	Almehmadi, Abdulaziz; Joudaki, Zeinab and Jalali, Roozbeh	2017	LDA	ACM DIGITAL LIBRARY
9	Data miners' little helper: data transformation activity cues for cluster analysis on document collections	Tania Cerquitelli; Evelina Di Corso; Francesco Ventura; Silvia Chiusano;	2017	PCA	ACM DIGITAL LIBRARY
10	Tea in Benefits of Health: A Literature Analysis Using Text Mining and Latent Dirichlet Allocation	CHING-HSUE CHENG; WEI-LUN HUNG;	2018	LDA	ACM DIGITAL LIBRARY
11	On a Topic Model for Sentences	Georgios Balikas; Massih-Reza Amini; Marianne Clause;	2016	LDA	ACM DIGITAL LIBRARY
12	Managing and Visualizing Citation Network Using Graph Database and LDA Model	Thuc Nguyen; Phuc Do;	2017	LDA	ACM DIGITAL LIBRARY

Fonte: Autoria Própria.

Os resultados da pesquisa por estudos primários que possibilitam comparações entre as técnicas LDA e PCA dentro da mineração de texto não foram satisfatórios. Devido a isso, foi realizada uma busca manual no *IEEE Xplore* com a *string* de busca "PCA" AND "LDA" que trouxe 876 resultados. Para aumentar o escopo dos resultados, foi necessária a retirada da *String* de busca dos termos *document*, *text* e *mining*. Desses resultados, foram lidos os títulos de todos, obedecendo aos critérios de inclusão e exclusão, em que foram escolhidos apenas 2. Foram considerados mais relevantes os artigos Martinez e Kak (2001), e Sushma Niket Borade e Adgaonkar (2011), forme Tabela 7.

Tabela 13 – Artigos sobre LDA encontrados e suas bases de dados.

Ordem	**Título**	**Autor**	**Ano**	**Técnica Abordada**	**Base de Dados**
1	PCA versus LDA	MARTINEZ, A.m.; KAK, A.c.	2001	PCA e LDA	IEEE Xplore
2	Comparative analysis of PCA and LDA	BORADE, Sushma Niket; ADGAONKAR, Ramesh P.	2011	PCA e LDA	IEEE Xplore

Fonte: Autoria Própria.

3.1.9 Discussão

A partir desta RSL e por meio das respostas que se obteve para este Capítulo para as questões de pesquisa da RSL do Capítulo Questões de Pesquisa 3.1.3, é possivel identificar que a aplicaçao do modelo estatístico LDA pode ser a mais apropriada para validação da ideia proposta neste trabalho em que os textos são dispostos consoante a gramática normativa, ou seja, trabalha-se com ambientes em que os textos são colocados em conformidade com as regras- sem gírias ou abreviações desnecessárias-, pois é um modelo estatístico que faz a descoberta de tópicos em uma coleção de textos, usado para descobrir semânticas que estão ocultadas em um dado corpo de texto. Além disso, trabalha com a correlação de termos, o que traz como conseguência resultados que se aproximam mais do contexto de uma população de termos, possibilitando inferir melhor sobre a semântica do texto analisado.

Por meio dessa RSL, foi possível identificar artigos que trazem como solução a aplicação dos métodos LDA e PCA sobre a extração de comportamento semântico em textos, e conseguiu-se também fazer a identificação de estudos experimentais que tiveram uma condução para a validação desses métodos. Não se conseguiu extrair, por meio desta revisão sistemática, qual modelo tem melhor acurácia quando aplicados ao mesmo conjunto de dados acerca do fenômeno mineração de texto devido a não existir algum trabalho que

faça a comparação entre essas técnicas.

Dessa forma, nesta seção, são apresentadas as respostas das questões de pesquisa definidas no protocolo desta RSL:

- **Questão primária (QP1)**: Como os métodos LDA e PCA podem ser utilizados para a extração de comportamento semanticamente em textos?

Para Balikas, Amini e Clausel (2016), o método LDA realiza o reconhecimento de padrão por meio da separabilidade de classes, enquanto que Cerquitelli *et al.* (2017) coloca o PCA não levando essa característica em consideração. Ainda em concordância com Martinez e Kak (2001), deixam bem claro que o LDA trabalha diretamente com a discriminação entre classes, enquanto o PCA trabalha com os dados em sua totalidade para a análise de componentes principais, sem prestar qualquer atenção particular à estrutura de classe subjacente. O método LDA pode ser utilizado para a extração de comportamento semântico em textos por meio de adaptações, pois em sua forma clássica pode ser utilizado em textos longos em que as palavras aparecem de modo formal.

- **Questão secundária (QS1)**: Qual estudo experimental conduzido para validar os métodos LDA e PCA?

A fim de responder à QS1, mostra-se que os trabalhos avaliados nesta RSL possuem peculiaridades que fizeram com que haja várias formas de validação:

  - Alhawarat e Hegazi (2018) realizaram a validação por meio do método combinado aplicado a vários conjuntos de dados árabes. Estes estão disponíveis gratuitamente na internet e usado para verificar e confirmar a exatidão do método combinado. Diferentes conjuntos de dados de notícias árabes são usados no estudo para validar a metodologia;
  - Zhang, Mao e Zeng (2016) realizaram uma maneira tradicional de avaliar modelos de tópicos comparando perplexidade ou verossimilhança marginal por meio de um teste;
  - Zeng, Cheung e Liu (2012) usaram perplexidade preditiva como uma função para validação do número de tópicos, por meio de uma validação cruzada em diferentes tópicos, onde a menor perplexidade indica a melhor capacidade de energização para o conjunto de teste invisível;
  - Bakhshaei, Safabakhsh e Khadivi (2019) realizaram a validação para a saída do modelo, em que foi considerado o efeito de dados extraídos sobre a qualidade de um sistema SMT (Statistical Machine Translation) existente;

- Em Cerquitelli *et al.* (2017), o primeiro desenvolvimento de SELF-DATA em execução no Apache Spark foi validado em 5 coleções de documentos;
- Cheng e Hung (2018) usam a perplexidade como uma métrica de avaliação no que é usada para medir a distribuição de probabilidade que também pode ser usada como um modelo de previsão. Um menor valor de perplexidade indica melhor desempenho de generalização;
- Balikas, Amini e Clausel (2016) realizaram experimentos para verificar a aplicabilidade e avaliar o desempenho de senLDA comparado ao LDA. O processo é dividido em duas etapas: (i) a fase de treinamento, onde os modelos de tópicos são treinados para aprender os seus parâmetros, e (ii) a fase de inferência que é para documentos novos, em que não são vistos suas distribuições de tópico.

  Esta RSL buscou, por meio da execução da *string* de busca em bases de dados reconhecidas da computação, a aplicação das técnicas LDA e PCA em artigos considerados relevantes. Não obstante, encontrar nesses artigos uma maneira de os comparar é uma tarefa desafiadora, pois não é trivial essa comparação na área de mineração de texto. Essa comparação não é imparcial, haja vista os artigos encontrados utilizarem bases de dados diferentes e são voltados para temas diferentes. Como a pesquisa não retornou comparações entre essas técnicas para análise textual, mas possiblitou saber que quando o conjunto de dados é pequeno, PCA tem um melhor desempenho do que o LDA, pois as componentes principais conseguem revelar as características principais; já para uma quantidade de dados muito grande, LDA apresenta melhor desempenho (Martinez; Kak, 2001) e (Sushma Niket Borade; Adgaonkar, 2011). Também é possível dizer com esta RSL que o PCA apresenta uma sensibilidade menor que o LDA quando se tem diferentes conjuntos de dados de treinamento (Martinez; Kak, 2001) e (Sushma Niket Borade; Adgaonkar, 2011), isso implica que caso se tenha um espaço amostral muito grande em uma conversação de um fórum, LDA pode retornar resultado mais próximos do esperado.

- **Questão secundária (QS2)**: Em termos de eficiência, qual método (LDA ou PCA) apresenta melhores resultados?

Esta RSL buscou a aplicabilidade acerca das técnicas LDA e PCA em artigos que são considerados relevantes e trouxe como resultado o fato de que comparar essas técnicas diretamente não é algo trivial no universo da mineração de texto, pois os artigos encontrados que trazem o PCA sendo comparados com LDA faz essa comparação no reconhecimento facial, como pode observar na Tabela 13. Essa comparação também não é imparcial, haja vista os artigos encontrados na mineração de textos utilizarem bases de dados diferentes e são voltados para temas diferentes.

Enquando no PCA os fatores ortogonais (componentes principais) não são correlacionados, no LDA há correlação entre os termos quando se quer reduzir a dimensionalidade dos dados. Assim, mesmo os resultados dessa pesquisa não serem realizados diretamente em cima de fóruns de discussão de um AVA, como a estrutura gramatical normativa é seguida, é possivel afirmar que para um ambiente em que os dados textuais estão dispostos de maneira formal os fatores ortogonais não são muito apropriados a para serem analisados, pois há uma certa perda de informação, enquanto as classes reportadas pelo LDA são mais coerentes com o contexto de um conjunto de dados e podem sim serem aplicadas a um fórum de discussão de acordo co aproposta deste trabalho de dissertação, pois os resultados denotam que existe uma proximidade semântica e, por conseguinte, do tema em discussão nesse ambiente.

3.1.10 Conclusões e Limitações da RSL

Por meio de uma ampla análise de métodos empíricos, que fazem observações práticas de análise, testes e comprovações, esta RSL busca capturar informações sobre os efeitos que as técnicas LDA e PCA tem sobre o fenômeno mineração de texto, obtendo resultados consistentes que fornecem evidências de que a utilização delas é robusta o suficiente. Dessa forma trazer respostas coerentes no universo da redução da dimensão de dados na forma textual para reconhecimento de padrões, de modo que, se possa comparar ambas as técnicas. Para as pesquisas realizadas nas bases de dados da Tabela 7, a quantidade de registros do método LDA dentro da área de mineração de texto é muito maior que a quantidade de registros do método PCA. Apenas duas ocorrências com o método PCA foram consideradas relevantes, como se pode ver na Tabela 12, o que trouxe como consequência a realização de uma busca manual no *IEEE Xplore*, conforme Tabela 13. Verifica-se que o PCA, quando se trabalha com pequenas quantidades de dados, apresenta melhor resultado que o LDA (CERQUITELLI *et al.*, 2017). Torna-se importante observar que esta RSL revela que a aplicação do método LDA sofre alterações que dependem do contexto quando se quer realizar o reconhecimento de padrões em cima do "*small text data*". Essas alterações são realizadas porque em textos curtos se tem poucos dados para se fazer extração de características relevantes, além disso alguns conjuntos de dados textuais trazem ruídos (informalidade e gírias). Segundo Zhang, Mao e Zeng (2016), textos curtos apresentam como característica negativa para se fazer análise textual porque são substancialmente diferentes dos documentos clássicos de texto, uma vez que são geralmente mais informal, mais ruidoso e menos focado no tópico. Zhang, Mao e Zeng (2016) também ressaltam que para textos curtos é bem mais difícil fazer a descoberta de padrão do que para grandes quantidades de textos. Essa dificuldade não diz respeito apenas à quantidade de textos, mas a ruídos também, pois uma quantidade pequena de texto instantâneo pode vir com gírias e abreviações que não fazem parte dos padrões do idioma.

3.2 Os Artigos Relacionados

A RSL possibilitou, entre outras coisas, a condução dos artigos relacionados.

3.2.1 Detecção Conjunta Fracamente Supervisionada do Sentimento-Tópico de Textos

Com o advento da *Web* 2.0, os usuário passaram não apenas a ler as informações, como também passaram a postar opiniões em diversos tipos de mídia, como *blogs*, fóruns de discussão e redes sociais. Isso apresenta uma riqueza de informações que podem ser útil para avaliar o sentimento do público em geral e opiniões sobre produtos e serviços. A pesquisa de Lin *et al.* (2012) revela que os recursos ricos em opiniões, como revisões *on-line*, estão tendo um impacto econômico maior em consumidores e empresas, em comparação com a mídia tradicional. Impulsionado pela demanda de coletar *insights* sobre grandes quantidades de dados gerados pelos usuários, há pesquisas que visam trabalhar em novas metodologias para análise automatizada de sentimentos e, também, descoberta de conhecimento oculto a partir de dados de texto não estruturados. Baseado nesse contexto, a pesquisa de Lin *et al.* (2012) faz uso de uma nova estrutura de modelagem probabilística chamada de JST (*Joint Sentiment Topic*), modelo baseado na LDA, que detecta sentimento e tópico simultaneamente a partir do texto (Lin *et al.*, 2012).

Dessa forma, esta pesquisa traz um diferencial, por usar também uma versão reparametrizada do JST chamado *Reverse-JST*, obtido pela reversão da sequência de sentimento e geração de tópico no processo de modelagem. O JST é uma escolha razoável de modelo para detecção de tópico de sentimento, sabendo-se que os sentimentos podem variar de acordo com tópicos (Lin *et al.*, 2012).

Como análise positiva desta abordagem, tem-se que o JST é um algoritmo razoável para análise de textos curtos, logo não é uma boa opção quando se quer trabalhar com documentos de textos muito grandes.

3.2.2 Um Modelo de Tópico não Paramétrico para Textos Curtos que Incorpora o Conhecimento de Coerência de Palavras

Há uma grande quantidade de textos curtos, como *tweets*, mensagens instantâneas e anúncios, que são difundidos no ambiente da *Web*. Há algum comportamento por trás desses textos em que um modelo de mineração de tópicos pode ajudar a entender os principais conteúdos implícitos e direcionar aplicativos de mineração, por exemplo, perfis de usuário, recomendação, difusão de informação e análise de influência. Para textos curtos é bem mais difícil fazer a descoberta de padrão do que para grandes quantidades de textos. Essa dificuldade não diz respeito apenas à quantidade de textos, mas a ruídos também,

pois uma quantidade pequena de texto instantâneo pode vir com gírias e abreviações que não fazem parte dos padrões do idioma (ZHANG; MAO; ZENG, 2016).

Com o objetivo de fazer uma análise textual numa grande quantidade de textos curtos, a pesquisa de Zhang, Mao e Zeng (2016) mostrada neste subtópico, propõe a utilização da técnica *Biterm Topic Model* (BTM), um modelo de padrões de co-ocorrência de palavras. Ele se revela eficaz para detecção de tópicos em textos curtos, mas suscetível a resultados limitados devido à escassez de textos e ruídos que neles aparecem.

3.2.3 Uso de Linguagem no Twitter Prevê Taxas de Criminalidade

A utilização de redes sociais produzem atualmente enorme quantidade de dados textuais. *Twitter* [1], que é uma rede de *microblog*, que contém mais de 230 milhões de usuários ativos, são postados mais de 500 milhões de *tweets* diariamente. O trabalho de Almehmadi, Joudaki e Jalali (2017) tem como objetivo analisar dados públicos do *Twitter* para prever taxas de criminalidade nas cidades por meio de termos que parecem ser ofensivos. Assim, faz-se uso da linguagem utilizada no *Twitter* de forma que haja a previsão das taxas de criminalidade e possibilita fazer uso da análise de *tweets* que foram coletados por um período de 3 meses em *Houston* e *Nova York*, bloqueando a coleção por longitude e latitude geográficas. Além disso, *tweets* sobre eventos criminais nas duas cidades foram coletados para verificação da validade do algoritmo de previsão. Foi utilizado o classificador *Support Vector Machine* (SVM) para criar um modelo de previsão de taxas de criminalidade baseado em *tweets*.

A proposta de Almehmadi, Joudaki e Jalali (2017) trabalha com a SVM como um classificador baseado em palavras previamente determinadas e não se apresenta tão eficaz quanto à automação de classificar os termos dentro de um contexto, sem se basear em um conjunto pré-definido de palavras, devido, muitas vezes, às palavras não se apresentarem no contexto literal.

3.2.4 Sobre os Trabalhos Relacionados

Os trabalhos relacionados mostrados propõem análise textual com base em textos curtos, onde se desenvolvem algoritmos que não são uma boa opção, quando se quer trabalhar com esse tipo de texto e existe a necessidade de alguma alteração ou limitação devido estar suscetível à utilização de termos não formais. Esses trabalhos revelam-se eficazes para detecção de tópicos em textos curtos, mas passível de resultados limitados devido à escassez de textos e ruídos que neles aparecem. Eles possuem também essa fragilidade, mas buscam resolver esse problema reduzindo a dispersão com enriquecendo dos modelos de representação de texto curto, que podem então ser manipulados por

[1] https://www.twitter.com/

métodos tradicionais de classificação, como mostrado por Flisar e Podgorelec (2018). A proposta de Almehmadi, Joudaki e Jalali (2017) visa à classificação de termos no documento através da SVM, como um classificador, baseado em palavras previamente determinadas, não sendo tão eficaz quanto à automação de classificar os termos dentro de um contexto sem se basear num conjunto pré-definido de palavras.

Assim, como a proposta de pesquisa e desenvolvimento deste estudo estão direcionados à análise de uma conversação na parte de fórum de discussão de um AVA, possibilitando o desenvolvimento de um módulo que trabalha investigando textos formais. Caso houvesse textos informais, haveria a necessidade da aplicação de técnicas com alguma alteração no algoritmo.

Ao se bsucar resultados consistentes em bases de dados reconhecidas da computação acerca das técnicas propostas, verificou-se que elas ainda não foram utilizadas em fóruns de um AVA para ajudar a área da educação. Sendo esta uma contribuição relevante e diferente que este trabalho pode deixar em releção aos outros trabalhos investigados.

4 Um Módulo de Mineração de Dados não Estruturados de Alunos em um Ambiente Virtual de Aprendizagem

Diante da intensificação de grupos de estudo em fóruns de discussão por meio de um AVA, num universo que possui muitos dados não estruturados, este estudo se apresenta como uma proposta de analisar se uma conversação textual está dentro de determinado objetivo de aprendizagem. A busca por essa análise ajuda a descobrir quais são os fatores que podem influenciar os alunos a desviarem o contexto proposto. Assim, é possível também analisar os textos a fim de que se conseguir ter os termos relevantes dando controle ao docente sobre o andamento do tema em discussão. Busca-se também analisar a eficácia da mineração de dados aplicada na mediação entre docente e alunos.

Uma das metodologias utilizadas para alcançar os objetivos deste trabalho é a utlização de questões de pesquisa, por meio do padrão *design science*, em que a construção e aplicação de um artefato trazem conhecimento, compreensão e solução para um conjnto-problema. Como se pode observar nesta seção, para responder a essas questões de pesquisa, foram necessários testes iniciais em livros, nos quais se obtiveram os resultados preliminares e possibilitaram se responder a algumas questões de pesquisa; e testes em bases de dados com textos de alunos que ajudaram a responder as demais questões de pesquisa.

4.1 RESULTADOS

Neste seção são respondidas questões de pesquisa por meio dos resultados considerados relevantes. Dentro dos resultados são encontrados os resultados preliminares utilizados, para validar os modelos estatísticos, e os resultados que trazem respostas em relação à técnica proposta de análise textual.

4.1.1 Resultados Preliminares

Para Wieringa (2014) para se obter a resolução da Questão Geral de Pesquisa definida na Introdução deste trabalho, é preciso seguir um conjunto de métodos que estão relacionados ao conjunto problema e ao conjunto solução. Dessa forma, é necessária a utilização dos métodos para que as questões presentes no conjunto problema sejam resolvidas, trazendo como resultado o conjunto solução. Como pode ser visto na Figura 2, uma QGP constitui o conjunto problema, em que ela é subdividida em questões QC, CT e

QP. Também se tem o conjunto solução, que são os artefatos gerados pelas respostas de cada questão, junto com os métodos utilizados.

Para este estudo e análise de resultados preliminares que ajuda a selecionar a técnica mais eficiente entre LDA e PCA, vamos utilizar a linguagem R, pois não é apenas uma linguagem de programação, mas um ambiente interativo amplamente utilizado na área de Ciência de Dados. Como recurso adicional, utilizou-se o ambiente de desenvolvimento integrado *RStudio*, que permite aproveitar algumas vantagens da linguagem R, tais como ser inteiramente gratuito e já vem com distribuições compiladas para *Windows*, *Mac* e *Linux* (OLIVEIRA; GUERRA; MCDONNELL, 2018), além disso apresenta uma documentação de fácil compreensão, simultaneamente *on-line* em diversos formatos (WICKHAM; GROLEMUND, 2016).

Nesta seção são apresentadas as respostas para as questões de pesquisa, que foram encontradas a partir da análise dos trabalhos selecionados e da aplicação dos modelos estátisticos.

A elucidação dos resultados preliminares começa baseando-se em responder a questão geral de pesquisa (QGP), proposta em 1.5, definida assim: **como analisar se um texto produzido por um grupo de estudo em um Ambiente Virtual de Aprendizagem está em conformidade com um determinado objetivo de aprendizagem?** A resposta direta a essa pergunta é por meio da mineração de texto. Como visto em RSL 3.1, a descoberta de conhecimento dentro de um determinado conjunto de dados não estruturados e textuais pode ser realizada por meio da utilização das técnicas LDA e PCA da estatística.

Quando os alunos utilizam o Ambiente Virtual de Aprendizagem, eles produzem dados que podem ser utilizados para ser analisados e inferidos. Técnicas da estatística permitem que se extraia os termos que são mais relevantes. Como visto em 1.6, é possível, por meio da metologia utilizada pela Ciência de Dados, utiliza recursos da estatística para, a partir da aplicação das técnicas LDA e PCA, coletar dados na forma textual, analisá-los e interpretá-los e entender como eles estão relacionados aos modelos propostos para o problema em estudo. Assim, ao comunicar os resultados da aplicação dos modelo propostos para análise de dados na forma textual, é possível gerar não apenas um tópico relacionado ao contexto em que os alunos estão debatendo, mas ainda coordenadas (componentes principais) no universo multidemensional do conteúdo produzido pelos alunos.

1. (QC) **Quais fatores podem influenciar alunos a desviarem o contexto proposto?**

   Resultados podem ser vistos em Questão Conceitual 3.1.9.

   a) Quais dados devem ser utilizados e como extraí-los para desvendar as causas do desvio ao tema proposto?

Resultados podem ser vistos em Questão Conceitual - (a) 1a

2. (QT) **Como se deve analisar dados na forma textual para que se consiga ter os termos mais relevantes dando controle ao docente sobre o andamento do tema proposto?**

   Em resposta a questão QT, este trabalho usa a Ciência de Dados e métodos da estatística para se fazer a redução da dimensão de dados na forma de texto. Embora os dados sejam apresentados na forma textual a partir do usuário, o computador utiliza números para se fazer qualquer cálculo. Dessa forma, o primeiro passo antes de inserir dados nos modelos estatísticos propostos para este trabalho é transformá-los em números por meio de uma matriz de termos de documentos. Assim, é possível explorar conceitos da estatística em aplicações de análise textual.

   a) Quais métodos considerados relevantes para a análise de dados na forma textual?

      A Ciência de Dados se preocupa inicialmente em quais dados são importantes e como extraí-los. Depois de decidido quais são o dados imprescindíveis para a análise, vem a etapa de transformação da Ciência de Dados que, por intermédio das técnicas propostas neste trabalho, aplicam-se os modelos estatísticos LDA e PCA, transformando dados em informacao e conhecimento de modo a possibilitar inferir sobre algum comportamento.

      A revisão sistemática, que pode ser vista RSL 3.1, mostra que é possível por meio dos métodos LDA e PCA da estatística que se pode reduzir a dimensão dos termos em cada documento selecionado deixando aqueles que mais são relevantes de acordo com a ideia implementada por trás de cada técnica. Cada método estudado oferece respostas peculiares, mesmo com a utilização da mesma base de dados. Dessa forma, é possível realizar um exploração de conceitos da estatística, por meio das técnicas LDA e PCA, em aplicações de análise textual.

      Os fóruns de discussão em AVA possuem como uma das características a utilização pelos usuários de frases construídas por meio do padrão nomativo do idioma. Logo, os modelos apresentados para essa fase de testes preliminares reportam resultados de acordo com a quandidade de informação formal postadas em abreviações e gírias. Para obtenção dos resultados preliminares, que ajudam a responder a questão de pesquisa, e também, possibilita trazer testes iniciais nas técnicas propostas, ajudando a selecionar a mais eficiente, é utilizado como padrão, para se fazer a análise textual ,obras do projeto Gutenberg. Este possui livos antigos considerados de domínio público, pois os direitos autorais deles nos EUA já expiraram. O projeto *Gutenberg* é uma biblioteca *on-line* de *eBooks* gratuitos e é o primeiro fornecedor de livros eletrônicos gratuitos. Dessa forma, esse pacote possui livros que podem ser acessados por um ID para a extração de caraterísticas. São escolhidos os livros *Twenty Thousand Leagues under the*

Sea, The War of the Worlds, Pride and Prejudice e Great Expectations desse pacote para se fazer a comparação preliminar entre as técnicas LDA e PCA.

b) Quais resultados da análise de dados textuais esses métodos reportam que podem ser considerados relevantes?

Como visto em 2.3.1, a técnica LDA possui como ideia central encontrar a combinação linear de recursos que melhor separa os dados na quantidade de classes desejadas, enquanto que em 2.3.2, tem-se a técnica PCA que tem o objetivo de encontrar os campos realmente importantes em bancos de dados com um grande número de variáveis, preservando a maior variabilidade possível

i. Resultados Preliminares do Modelo LDA

Sabendo que a modelagem de tópico é uma abordagem baseada em probabilidade para encontrar *clusters* dentro de documentos (KWARTLER, 2017) e que para o pacote *utenberg*, cada livro pode ser tratado como uma mistura de tópicos e cada tópico como uma mistura de palavras, nesta seção, portanto, mostram-se os resultados de tópicos como uma mistura de palavras do método lda() para os livros *Thousand Leagues under the Sea, The War of the Worlds, Prideand Prejudice e Great Expectations*, haja vista esse modo se assemelhar mais a contexto de temas, ou seja, o tema em torno do qual gira o contexto.

Tendo em vista que se está trabalhando com os livros *Thousand Leagues under the Sea, The War of the Worlds, Prideand Prejudice e Great Expectations*, usa-se o modelo LDA, por meio da função *LDA()* da linguagem R, para criar um modelo com quatro tópicos buscando a relevância dos termos. A Tabela 14 mostra as probabilidades por tópico por palavra dos quatro primeiros termos dos tópicos relacionados aos livros *Twenty Thousand Leagues under the Sea*, *The War of the Worlds*, *Pride and Prejudice* e *Great Expectations*.

Consegue-se extrair, como visto na Tabela 14, tópicos com os termos *joe*, *biddy*, *estella* e *pocket*. Para cada combinação, o modelo LDA calcula a probabilidade desse termo está sendo gerado a partir de um tópico. Por exemplo, o termo "*joe*" tem uma probabilidade quase zero de ser gerado a partir dos tópicos 1, 2 ou 3, mas representa 1,45% do tópico 4.

Como se pode observar na Tabela 15, os cinco principais termos dentro do tópico 2, relacionados ao livro *Twenty Thousand Leagues under the Sea*, são:

Tabela 14 – Probabilidades por tópico por palavra.

Tópico	Termo	Beta
1	joe	1.436612e-17
2	joe	5.962111e-61
3	joe	9.881855e-25
4	joe	1.447329e-02
1	biddy	5.139275e-28
2	biddy	5.022015e-73
3	biddy	4.307280e-48
4	biddy	4.775557e-03
1	estella	2.431464e-06
2	estella	4.323253e-68
3	estella	1.552457e-18
4	estella	4.961910e-03
1	pocket	2.210970e-04
2	pocket	2.519701e-05
3	pocket	1.291206e-04
4	pocket	3.799413e-03

Fonte: Autoria Própria

Tabela 15 – Cinco principais termos do tópico 2.

Tópico	**Termo**	**Beta**
2	captain	0.015510635
2	nautilus	0.013051927
2	sea	0.008843483
2	nemo	0.008709651
2	ned	0.008031955

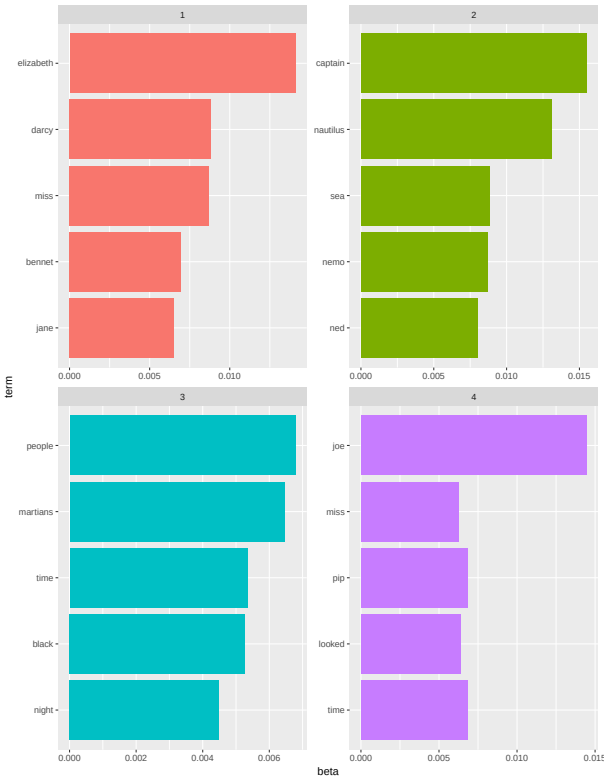
Fonte: Autoria Própria

O modelo LDA possui uma clusterização difusa que implica que um termo pode ter probabilidades de pertencer a mais de um tópico. Isso pode ser visto na Figura 13 em que se consegue ver que o termo *miss* pertence tanto ao tópico 1 quanto ao tópico 4, e o termo *time* possui também a probabilidade de pertencer a mais de um tópico, no caso ao tópico 3 e ao tópico 4.

ii. Resultados Preliminares do Modelo PCA

Para a obtenção de resultados preliminares para o modelo PCA, nesta seção, buscam-se encontrar os campos que realmente são importantes no

Figura 13 – Vizualização gráfica dos livros analisados.



Fonte: Autoria Própria

textos. Para essa tarefa, usa-se a Análise de Componentes para encontrar os componentes principais de dados. Assim, com a ajuda de transformações ortogonais, este trabalho propõe a utilização do PCA com o objetivo de interpretar os dados em uma estrutura mais significativa. Essas transformações destinam-se a mostrar a estrutura interna do conjunto de dados com uma nova base arbitrariamente projetada no espaço vetorial, o que melhor explica a variação dos dados, isto é, calculam-se novas variáveis a partir dos dados originais, onde estas novas variáveis incluem a variância das variáveis originais em ordem decrescente (DARÓCZI, 2015). De qualquer forma, o PCA pode enviar com sucesso uma projeção de menor dimensão dos dados, onde as componentes principais não correlacionadas são as combinações lineares das variáveis originais. Com essa visão geral informativa, pode-se ter uma grande ajuda para a análise quando se identifica a estrutura subjacente das variáveis (DARÓCZI, 2015).

A partir de um conjunto de variáveis originais, o método PCA extrai um número de componentes principais. É importante observar que o primeiro componente inclui o que possui maior variação comum, por isso tem a maior importância na descrição do conjunto de dados originais; enquanto o último muitas vezes inclui apenas algumas informações exclusivas de somente uma variável original (DARÓCZI, 2015).

A Figura 14 mostra 10 componentes principais para o livro *Twenty Thousand Leagues Under the Sea*. Em cada célula, entre uma palavra e o termo PC_n, há um valor que indica a probabilidade de o termo da célula ser uma comopnente princial. Observa-se que o termo *frigate* possui maior possibilidade de ser a primeira componente principal (PC1) com 6,5%, sendo esse percentual maior que os dos demais termos disponíveis para PC1. As outras componentes principais recebem as seguintes probabilidades em relação às demais candidatas:

A. PC2, tem-se ned com 43%;
B. PC3, tem-se também ned com 36%;
C. PC4, tem-se captain com 25%;
D. PC5, tem-se sir com 19%;
E. PC6, tem-se nautilus com 17%;
F. PC7, tem-se conseil com 32%;
G. PC8, tem-se air com 8,8%;
H. PC9, tem-se conseil com 15%;
I. PC10, tem-se air com 14,9%.

Figura 14 – 10 componentes principais.

	PC1	PC2	PC3	PC4	PC5
captain	-5.409760e-01	1.606030e-01	-1.805092e-01	2.557119e-01	1.222147e-01
nautilus	-4.621065e-01	-2.248340e-01	2.874795e-01	-2.735808e-01	-1.031774e-01
sea	-1.761080e-01	-4.733058e-03	-2.504327e-01	-3.179315e-01	-1.916476e-01
ned	-2.696773e-02	4.333228e-01	3.613295e-01	-1.250184e-01	1.415979e-02
ice	-1.629427e-01	-2.033235e-01	1.648061e-01	2.864607e-02	7.475023e-02
water	-1.088410e-01	-1.913857e-01	2.043016e-01	-6.159897e-03	1.535637e-01
sir	-1.468679e-01	1.073939e-01	-1.035157e-01	-1.850389e-01	1.955969e-01
red	-5.450952e-02	2.127218e-02	-8.532922e-02	-1.864896e-01	-1.720427e-01
conseil	3.730830e-02	2.157659e-01	1.938121e-01	9.275695e-02	-9.084101e-02
nemo	-3.471342e-01	8.968985e-02	-1.279772e-01	1.504421e-01	3.530166e-03
air	-7.587482e-02	-8.474652e-02	1.541224e-01	5.578609e-02	6.810663e-02
land	-5.812528e-02	3.088312e-01	1.260827e-01	-5.501514e-02	-8.421789e-02
vessel	-6.195852e-02	-2.079932e-02	-2.476540e-02	-1.076145e-01	8.536647e-02
whale	-1.012538e-03	5.593960e-02	5.013152e-02	-3.780156e-02	1.226912e-04
frigate	6.569567e-02	4.511864e-02	-2.729391e-02	-4.255799e-02	1.401181e-01
vanikoro	2.405437e-03	-2.709489e-02	-5.034008e-02	3.870571e-02	-4.133810e-02
pressure	2.748933e-02	-4.055592e-02	5.489751e-02	-1.853566e-02	8.929006e-02
sun	-2.747307e-02	-9.145739e-03	-2.837135e-02	1.096477e-01	-8.749091e-02
professor	-9.454019e-03	1.610299e-02	-8.609096e-02	-5.010137e-02	1.596413e-01
savages	-3.242895e-02	2.221362e-02	1.190470e-03	5.320500e-02	2.933029e-03
fish	-4.063476e-03	-6.580242e-04	-3.062434e-02	-4.164411e-02	-7.126287e-02

	PC6	PC7	PC8	PC9	PC10
captain	9.659535e-02	-7.836160e-02	-1.043256e-01	-3.846081e-02	6.519403e-02
nautilus	1.700058e-01	-2.297407e-02	2.871568e-02	1.090548e-01	-1.040844e-01
sea	-1.993051e-01	-3.201317e-02	7.255933e-02	-1.242308e-01	6.823752e-02
ned	4.917174e-02	-2.190528e-02	-2.532645e-02	-1.165229e-01	-1.282664e-01
ice	-2.452989e-01	1.801991e-01	-5.488814e-02	6.412768e-03	-2.075825e-02
water	2.983534e-02	-8.972075e-02	4.222860e-02	-2.300245e-01	6.549922e-02
sir	-1.722534e-01	7.390193e-02	-1.526945e-01	1.010624e-01	2.807343e-03
red	-1.230550e-02	-3.470386e-02	2.683718e-02	-1.002716e-01	1.380797e-01
conseil	-1.161225e-01	3.252750e-01	3.980438e-02	1.516342e-01	1.343325e-01
nemo	7.901519e-02	2.082611e-02	4.523706e-02	7.616665e-02	1.322196e-02
air	-1.823256e-02	5.304803e-02	8.821792e-02	-1.112997e-01	1.492701e-01
land	-3.603252e-02	-2.884993e-02	-1.951240e-02	-6.173367e-02	-1.596137e-02
vessel	9.539393e-02	-1.998607e-02	-7.166196e-02	1.320851e-02	-1.658766e-01
whale	-6.223318e-02	-1.389105e-02	-9.727662e-02	5.046361e-03	-8.547935e-02
frigate	-1.269048e-01	-7.672648e-02	-1.940089e-01	-4.925579e-02	1.121092e-01
vanikoro	1.215812e-01	8.613859e-02	-1.499931e-01	-1.319536e-01	-9.461801e-02
pressure	-1.432401e-03	-4.609115e-02	-1.938057e-03	-1.096267e-01	-5.931309e-02
sun	-1.437635e-01	4.385944e-02	-1.035562e-02	-2.230130e-02	-3.329102e-02
professor	1.437246e-02	1.106898e-01	6.863598e-02	-5.255553e-02	-3.031727e-02
savages	6.390813e-02	5.626236e-02	-8.383326e-02	3.993446e-02	8.286081e-02
fish	2.296093e-02	2.476891e-02	8.618866e-02	-6.489602e-02	4.346849e-03

Fonte: RStudio (2018).

Consegue-se extrair, como visto na Figura 14, que as componentes principais, como *frigate* e *captain*, do livro *Twenty Thousand Leagues Under the Sea* são termos associados ao título do livro.

3. (QP) **Como analisar o efeito e a eficácia da mineração de texto aplicada na mediação entre docente e alunos?**

   a) Quais os resultados da análise de desempenho, em termos de acurácia, ao aplicar as técnicas propostas no mesmo conjunto de dados?

   Para alcançar os objetivos, em um primeiro momento, por meio da seleção dos livros, o método LDA trouxe tópicos associados a cada um deles, em que foi possível visualizar como os termos de cada tópico é distribuído entre eles. E no segundo momento, após o método LDA retornar resultados dos quatro livros, foi escolhido um deles para o método PCA. Poder-se-ia escolher qualquer outro livro para a mesma análise. Esse método trouxe os termos que mais carregam informações dentro do espaço vetorial de dados do livro selecionado. Na aplicação do método LDA, nota-se que os quatro tópicos estão claramente associados aos quatro livros, pois o tema "*captain*", "*nautilus*", "*sea*", e "*nemo*" pertencem a *Twenty Thousand Leagues Under the Sea*, e que "*jane*", "*darcy*", e "*elizabeth*" pertence a *Pride and Prejudice*. Percebe-se que "*pip*" e "*joe*" são de *Great Expectations* e "*martians*", "*black*", e "*night*" de *The War of the Worlds*. Também se pode notar que, como o LDA é um método de clusterização difusa, pode haver palavras em comum entre vários tópicos, como "*miss*" nos tópicos 1 e 4, e "*time*" nos tópicos 3 e 4 (Figura 13), em que cada elemento em comum pertence a um grupo segundo uma probibilidade. Porém o método LDA, ao mostrar os percentuais de um termo pertencer a um determinado tópico, trouxe resultados inferiores quando comparados com o método PCA. Um exemplo é o termo *captain* que o LDA indicou 1,5% de pertencer ao tópico do livro *Twenty Thousand Leagues Under the Sea*, enquanto o PCA indicou 25% de aquele mesmo termo ser uma componente principal associada a esse mesmo livro. O termo *ned* tem 0,08% de pertencer ao tópico desse livro, mas possui 43% de ser uma componente principal. O termo *nautilus* possui 1,3% de fazer parte do tópico desse livro, enquanto como componente principal possui 17%. Com esses dados, pode-se concluir que o PCA apresenta, em termos quantitativos, resultados maiores, mas isso se deve a como o algoritmo reporta seus resultados. Em outras palavras, ambas as técnicas conseguiram trazer resultatos aceitáveis, porém o LDA trabalha matematicamente com o agrupamento dos termos mais próximos e que traz mais significado aos resultados. Desse modo, em termos de eficiência, em que independe o tamanho do percentual reportado por cada técnica, ambos os algoritmos apresentaram resultados similares. Como exemplo,

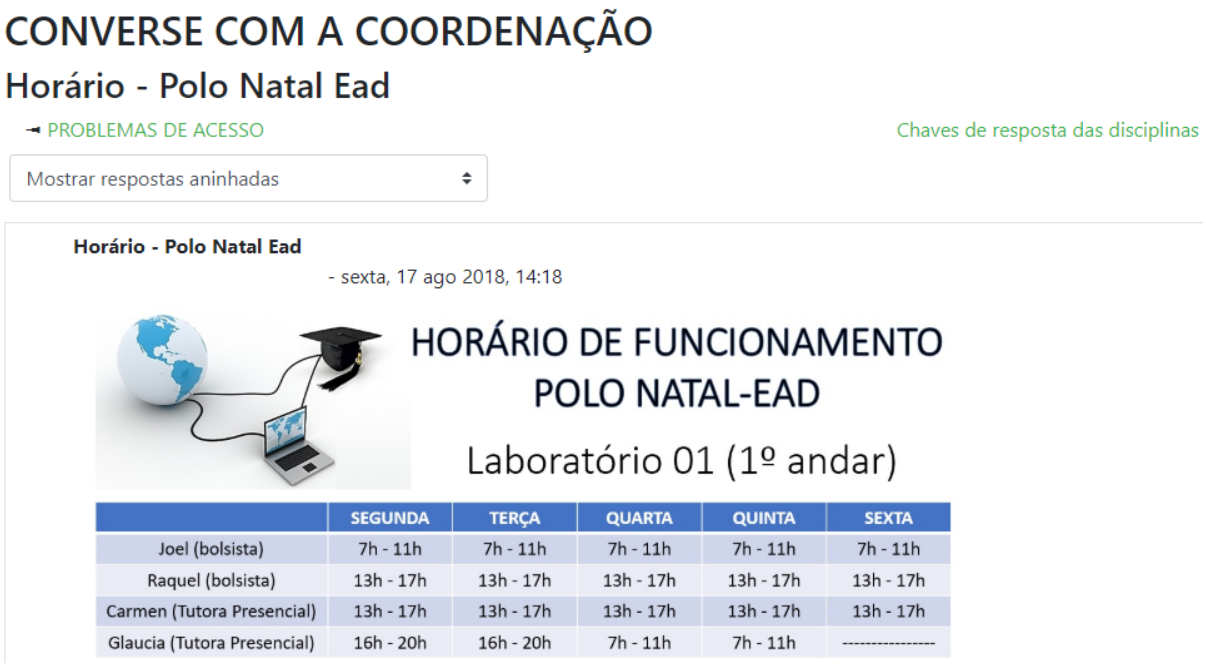
pode-se mostrar o termo *captain* que o LDA indicou 1,5%, enquanto o PCA, 25% em relação ao livro *Twenty Thousand Leagues Under the Sea*. A quantidade em percentual não indica necessariamente que um algoritmo aprensetou melhor acurácia que o outro, pois isso está intrissecamente relacionado aos cálculos utilizados internamente nos algoritmos. Porém, o modelo LDA trabalha com resultados que evidencia a correlação semântica entre os termos, o que se apresenta como uma melhor técnica quando o assunto é extrair um tema em torno do qual um texto se situa. Quando o algoritmo LDA retornou os termos mais relevantes, esses termos estão próximos lógico-matematicamente e por meio deles é possíve inferir o tema em tornou do qual o livro trabalha. Logo, por meio desses resultados, é possivel testar o LDA para se verificar se apresenta também resultados parecidos quando utilizado na ferramenta de fórum de um AVA, para se descobrir uma estrutura temática na qual os alunos discutem. Mesmo aplicando os algoritmos em livro como uma das formas de validar o algoritmo LDA, a inferência da estrutura temática que o LDA traz, pode ser testada em qualquer conjunto de textos bem esturturados para se descobrir o assunto em torno do qual é tratado.

4.1.2 Resultados na Base De Dados do AVA

Esta seção visa responder as questões de pesquisa, utilzando apenas o modelo estatístico selecionado como mais adequado para o contexto de um AVA, com dados coletados numa base real. A base de dados utilizada para obter os dados a serem analisados é do curso de Informática para Internet do IFRN, Campus Avançado de Natal- Zona Leste.

Dentro do Canal de Comunicação com a Coordenação, na opção de CONVERSE COM A COORDENAÇÃO, o tópico lançado no dia 17 de Agosto de 2018, pela professora, sob o título **Horário - Polo Natal Ead**, conforme a Figura 15, possui certa relevância devido à quantidade de *feedbacks*, que totaliza 25 iterações, e por isso foi selecionado para a primeia análise dos dados textuais.

Figura 15 – Página do AVA onde os foram extraídos alguns dados.



	SEGUNDA	TERÇA	QUARTA	QUINTA	SEXTA
Joel (bolsista)	7h - 11h	7h - 11h	7h - 11h	7h - 11h	7h - 11h
Raquel (bolsista)	13h - 17h	13h - 17h	13h - 17h	13h - 17h	13h - 17h
Carmen (Tutora Presencial)	13h - 17h	13h - 17h	13h - 17h	13h - 17h	13h - 17h
Glaucia (Tutora Presencial)	16h - 20h	16h - 20h	7h - 11h	7h - 11h	----------

Fonte: https://ead.ifrn.edu.br/ava/academico/mod/forum/discuss.php?d=49009.

Ao se aplicar o método LDA na conversação do tópico do portal EAD do IFRN com tema Horário - Polo Natal Ead, obtém-se como resultado termos que o algoritmo considera como importantes para o contexto abordado durante a discussão. Observa-se que os termos dizem respeito a um tema (discussão) que gira em torno da obrigatoriedade de os alunos terem que cumprir carga horária com a participação presencial no laboratório do polo para participarem de atividades do curso. Assim, Tabela 16 elenca os termos que prioritariamente fazem parte de toda discusão.

Tabela 16 – Resultado do modelo LDA ao fórum EAD como tema Horário - Polo Natal Ead.

Ordem	**Termo**	**Relevância**
1	presencialmente	2.15%
2	laboratorio	2.15%
3	polo	2.15%
4	voce	2.15%
5	atividades	2.15%
6	curso	1.70 %
7	disciplinas	1.47%
8	ifrn	1.25%
9	natal	1.25%
10	carga	1.02%
11	horaria	1.02%

Fonte: Autoria Própria.

É importante lembrar que um termo pode ser relevante pela sua frequência, visto

que, quanto mais vezes um termo aparecer, mais importante ele se torna para o texto, como se pode observar na Tabela 17, com o termo *polo* que não seria uma palavra que mereceria estar no topo da discussão apropriada no tópico Horário - Polo Natal Ead ao se analisar o contexto. Nesse sentido, um termo também pode ter sua relevância por meio de um outro algoritmo que não considera necessariamente a frequência, como é o caso do modelo LDA, que de acordo com a correlação entre os termos, tem-se aqueles que podem fazer parte de um tópico, como na Tabela 16, em que os termos *presencial* e *laboratório* condizem com a discussão em análise.

Tabela 17 – Resultado por frequência dos termos ao fórum EAD com tema Horário - Polo Natal Ead.

Ordem	**Termo**	**Relevância**
1	polo	0.02710843373493976
2	atividades	0.02710843373493976
3	presencial	0.024096385542168676
4	curso	0.02108433734939759
5	laboratorio	0.015060240963855422
6	Natal	0.015060240963855422
7	disciplinas	0.015060240963855422
8	carga	0.012048192771084338
9	horaria	0.012048192771084338
10	informatica	0.012048192771084338
11	ifrn	0.012048192771084338

Fonte: Autoria Própria.

É importante notar que quando o professor/tutor lança um tópico para discussão, ele próprio tem a capacidade de saber quais seriam os termos relevantes na discussão e que, imprescindivelmente, estariam no texto. Isso possibilita que o professor infira, após a aplicação de um modelo estatístico/probabilístico, o quão há desvio ao objetivo da discussão, possiblitando, pois, que de alguma forma haja uma notificação aos alunos para participarem ou não desviarem do tema proposto para aquele bloco.

Outro tópico neste curso que se apresenta como segunda posição de relevância é postado por um aluno no dia 26 de março de 2020 sob o tópico **Disciplina duplicada**.

Buscando a correlação entre os termos a partir do modelo LDA, a fim de extrair os termos com maior importância da discussão proposta pelo aluno Mykael, tem-se como resultado os dados mostrados na Tabela 18.

Tabela 18 – Resultado do modelo LDA ao fórum EAD como tema Disciplina duplicada.

Ordem	**Termo**	**Relevância**
1	disciplina	4.41%
2	orientacao	3.23%
3	seminario	2.65%
4	pratica	2.65%
5	profissional	2.65%
6	orientador	2.65 %
7	pagina	2.65%
8	projeto	2.06%
9	integrador	2.06%
10	tcc	2.06%
11	moodle	2.06 %

Fonte: Autoria Própria.

A importância que os termos possuem podem ser mostradas, dependendo da quatidade de informação textual, também pela técnia TF. Dessa forma, a Tabela 19 evidencia quais termos são considerados os mais relevantes quando se fala em frequência com que eles aparecem na discussão.

Tabela 19 – Resultado por frequência dos termos ao fórum EAD com tema Disciplina duplicada.

Ordem	**Termo**	**Relevância**
1	disciplina	0.05426356589147287
2	orientacao	0.03875968992248062
3	seminário	0.031007751937984496
4	pratica	0.031007751937984496
5	profissional	0.031007751937984496
6	orientador	0.031007751937984496
7	pagina	0.031007751937984496
8	projeto	0.023255813953488372
9	integrador	0.023255813953488372
10	tcc	0.023255813953488372
11	curso	0.023255813953488372

Fonte: Autoria Própria.

Observando as tabelas Tabela 18 e Tabela 19 é notado que o modelo LDA trouxe os termos relevantes parecidos com os termos trazidos pela técnica TF, em que a exceção fica apenas nos termos *moodle* e *curso* na possição 11ª em cada tabela. Por conseguinte, para uma pequena quantidade de textos, o modelo LDA não se diferencia muito de uma técnica simples como a TF. Assim, LDA traz melhores resultados para grandes quantidades de textos.

A Tabela 20 mostra os resultados da aplicação do modelo LDA ao tentar classificar os 11 primeiros termos das duas discussões. Para isso, essas discussões foram colocadas

concatenadas, ou seja, num mesmo arquivo foi colocada primeiro uma discussão e, em seguida, a outra. O algoritmo LDA rodou em todo texto das duas discussões e trouxe os resultados apresentados na Tabela 20.

Tabela 20 – Resultado do modelo LDA para classificar os termos por tópicos das duas discussões.

Tópico 1	Tópico 2
disciplinas: 1.71%	disciplinas: 1.95%
laboratorio: 1.31%	voce: 1.37%
presencialmente: 1.28%	presencialmente: 1.35%
voce: 1.27%	laboratorio: 1.32%
sim: 1.21%	curso: 1.2%
curso: 1.18%	sim: 1.17%
horarios: 0.96%	horarios: 0.91%
ifrn: 0.80%	natal: 0.83%
professores: 0.80%	ifrn: 0.79%
orientacao: 0.80%	professores: 0.79%
natal: 0.77%	orientacao: 0.77%

Fonte: Autoria Própria.

Ao analisar os resultados reportados pelo modelo LDA da Tabela 20, observa-se que tanto o tópico 1 como o tópico 2 possuem os mesmos resultados. Porém é possível observar que, nesse caso, o modelo fez a classificação por percentual, sendo que, para dois termo iguais, o de percentual maior tem que aparecer no seu respectivo tópico, haja vista o modelo LDA mostrar não apenas a probabilidade de um termo pertencer a um determinado tópico, como também, de um tópico pertencer a um determinado documento (ou conversação). Nesse contexto, é possível construir uma matriz de confusão para avaliar o modelo de classificação reportado. Para Casella, Fienberg e Olkin (2013), a matriz de confusão ajuda porque os elementos na diagonal principal da matriz representam indivíduos cujos status padrão foram previstos corretamente, enquanto que elementos fora dessa diagonal representam indivíduos que foram classificados incorretamente. Para isso, vamos usar as seguintes terminologias:

- Verdadeiro positivo: true positive — TP;
- Falso positivo: false positive — FP;
- Falso verdadeiro: true negative — TN;
- Falso negativo: false negative — FN.

A tabela que se pode gerar para termos a matriz de confusão segue a categorização da Tabela 21.

Tabela 21 – Categorizando as células da tabela.

	Tópico 1	Tópico 2
Tópico 1	TP	FP
Tópico 2	FN	TN

Fonte: Autoria Própria.

Após categorizar, é hora de inserir os dados de acertos fornecidos pela Tabela 20. Assim, temos a Tabela 22 com a matriz de confusão usando os dados que temos.

Tabela 22 – Quantificando as células da matriz de confusão.

	Tópico 1	Tópico 2
Tópico 1	9	1
Tópico 2	0	1

Fonte: Autoria Própria.

Olhando para os dados que temos em Tabela 22, o modelo previu 9 vezes corretamente termos como pertencentes ao tópico 1 e uma vez corretamente um termo pertencendo ao tópico 2. Um termo foi classificado incorretamente e este foi colocado na célula *FP*.

Em relação à acurácia, que nos diz quanto esse modelo acertou das previsões possíveis, no nosso contexto, nosso modelo teve uma acurácia de 91% (0.909091), haja vista ter acertado 10 das 11 previsões. A acurácia pode ser obtida pela razão entre o somatório das previsões corretas dividido pelo somatório das previsões, conforme a Equação 4.1.

$$accurary = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

Também é possível, a partir da matriz de confusão, extrair a precisão a fim de saber a proporção de identificações que foram positivas e que foram realmente corretas. Neste caso, se tem uma precisão de 90%, de acordo com a Equação 4.2.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Neste ponto deste trabalho de dissertação, é possível responder as questões conceituais de pesquisa supracitadas na Subseção 1.5 e que não foi possível responder por meio dos resultados preliminares.

1. (QC) **Quais fatores podem influenciar alunos a desviarem o contexto proposto?**

   A partir do tópico lançado no dia 17 de Agosto de 2018, sob o título **Horário - Polo Natal Ead**, conforme a Figura 15 e que obteve 25 interações, consegue-se

notar que a discussão é direcionada ao cumprimento de 4 horas semanais presenciais no polo por alunos do curso de Informática para Internet. Dessa maneira, é possíve observar que a ausência de três pontos são impactantes para o desvio ou ausência dos alunos na discussão: interesse, desenvolvimento do tópico e relevância.

No tocante ao interesse, tem-se um ponto bastante pessoal do aluno, pois os alunos que realmente estão envolvidos no curso e que demonstram realmente empenho são os que participaram da discussão a fim de tirarem dúvidas.

No que concerne ao desenvolvimento do tópico, o desenvolvimento bem elaborado pela professora e sem dualidade de sentido também possibilitou que os alunos prestarem atenção à discussão e se envolvessem nela.

Acerca da relevância, a informação importante que foi postada possibilitou também que houvesse uma quantidade maior de interações entre os alunos, haja vista o curso ser à distância e se ter a necessidade do cumprimento de carga horária semanal presencial.

a) Quais dados devem ser utilizados e como extraí-los para desvendar as causas do desvio ao tema proposto?

   Ao analisar como funciona um AVA, percebe-se que entre as várias opções de se buscar interagir com os alunos a que melhor se encaixa para se retirar termos relevantes, e que possibilite inferir algo, são os fóruns de discussão. A partir de um fórum desses, consegue-se instigar os alunos a pensarem e compartilhar seus pensamentos. Dentro do fórum, os dados relevantes são apenas os textos da discussão, tendo muitos *stop words* adicionais inerentes à plataforma, como os nomes dos participantes, data/hora, textos da própria página como o título, "caminho de rato"característico do *design web*, entre outros.

   Ao observar o AVA, verifica-se que a maneira de extrair esses dados é acessando à base de dados diretamente, a fim de se ter os textos da discussão e se aplicar o algoritmo desejado para este trabalho de dissertação.

4.1.3 Módulo de Mineração de Texto

Os fóruns de discussão são a melhor forma de se buscar inferir os termos mais relevantes numa base de dados de um AVA. A partir desses fóruns, consegue-se incentivar os membros a participarem do compartilhamento de ideias em forma de texto. Além disso, nos fórum os dados que são considerados relevantes são os textos da discussão. Para este trabalho de dissertação, foi necessário o desenvolvimento de um ambiente claro e objetivo que represente um fórum de discussão para se aplicar a mineração de texto com a técnica escolhida após pesquisas e testes.

4.1.3.1 Desenvolvimento

Ao observar o AVA, verifica-se que a maneira de extrair esses dados é acessando à base de dados diretamente, a fim de se ter os textos da discussão e se aplicar o algoritmo desejado.

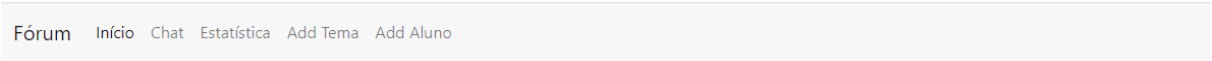
Com o objetivo de analisar se os alunos estão dentro do tema proposto, esta seção visa não apenas analisar o quão eficiente o modelo LDA é em um AVA, mas ainda de se obter *feedback* de aceitação por parte do docente (um especialista), e para isso foi necessária a implementação de um simples ambiente virtual em forma de fórum de discussão para que se consiga fazer a aplicação da técnica.

O ambiente implementado não traz todas as caracteríscas de um AVA, como datas de provas e trabalhos, materiais de estudo, entre outros, pois tem como objetivo analisar se os alunos se mantêm na temática proposta por meio de uma análise textual em uma discussão em forma de grupo de aprendizagem.

Para o desenvolvimento do módulo, foi utilizada a linguagem *PHP* 5.4, o *framework Bootstrap v4-alpha*, o *framework PHP PDO*, *xhtml*, *css*, *javascript* e *Google Chart Tools*. Todos os parâmetro enviados por formulários são passados por meio do método *POST*.

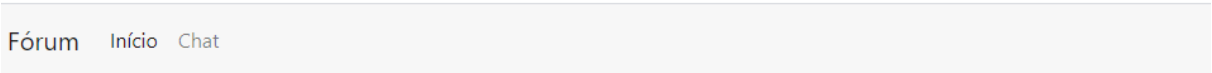
Há dois tipos de usuários para o módulo: o professor e o aluno. A Figura 16 mostra que o professor tem permissão para ver a estatística em torno do tema, adicionar os alunos que irão discutir e adicionar também o tema da discussão. Ao adicionar um aluno, este receberá na sua caixa de e-mail uma mensagem informando-o sobre a existência do fórum com seus dados de acesso. Para isso, o professor terá que informar o e-mail do aluno na hora do cadastro dele. As permissões do aluno podem ser vista na Figura 17.

Figura 16 – Permissões dadas ao professor.



Fonte: https://bit.ly/37lMaiL.

Figura 17 – Permissões dadas ao aluno.



Fonte: https://bit.ly/37lMaiL.

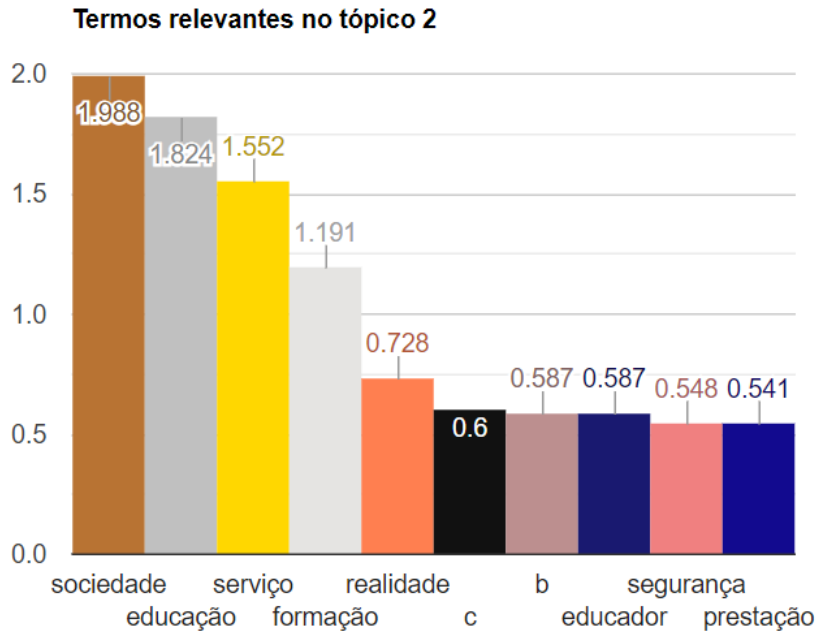
A ciência de dados visa saber quais e como os dados devem ser coletados, preocupando-se também em saber qual a melhor técnica de transformão de dados mais apropriada, além de buscar uma maneira de vizualizar as informações úteis geradas.

Ao se fazer a análise de dados, portanto, primeiro se coleta para se fazer um exame nos dados para apresentá-los. Baseando-se nas informações úteis capturas, busca-se a

visualização dessa informação a fim de otimizar a comunicação e a interpretação por meio de dados.

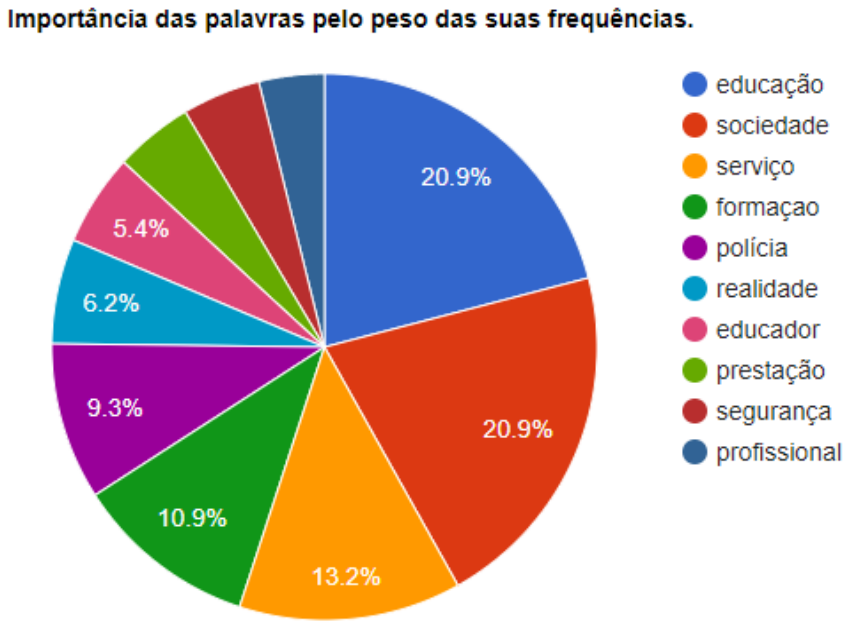
O ambiente vitural em forma de fórum desenvolvido aqui traz gráficos que auxiliam na interpretação dos dados gerados pelas técnicas aplicadas, ajudando o docente a entender qual o tema relevante que está sendo debatido pelos alunos. No ambiente desenvolvido, como se pode observar na Figura 16, supracitada, existe a opção estatística em que é possível a visualização dos termos mais relevantes por meio de gráficos. Por intermédio do gráfico de colunas, a Figura 18 mostra a relevância dos termos que o LDA traz como *feedback* para fazer parte dos tópicos que podem representar a conversa dos textos dos alunos, mostrando também, como exemplo, termos *stop words*. A Figura 19 mostra uma representação dos dados por meio do gráfico de pizza, em que quando se seleciona uma fatia, mostra-se a quantidade de termos com seus respectivos percentuais. Esses dois gráficos são exemplo de formas que este trabalho buscou para mostrar os resultados da análise textual.

Figura 18 – Visualização do LDA por meio do gráfico de colunas.



Fonte: https://bit.ly/37lMaiL.

Figura 19 – Visualização do TF por meio do gráfico de pizza.

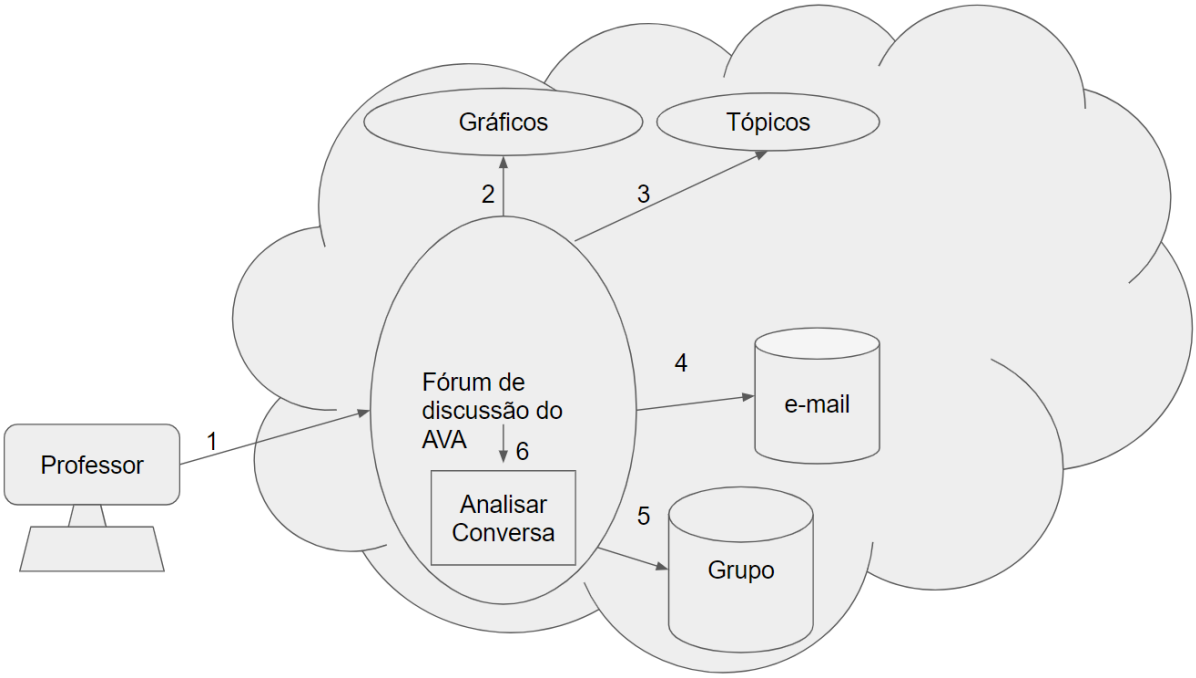


Fonte: https://bit.ly/37lMaiL.

A Figura 1 na *Introdução* mostra um modelo de Arquitetura ideal para a utilização a mineração de texto no AVA. A atuação de um Agente torna o sistema mais lento e como não é o foco deste trabalho, na prática temos apenas as funcionalidades do fórum de discussão. A Figura 20 mostra o modelo de Arquitetura que representa o ambiente desenvolvido para este trabalho de dissertação. Esse modelo mostra uma visão geral dos elementos que constituem a arquitetura geral do sistema. O fluxo do sistema, sob a ótica do professor, ocorre de acordo com os passos abaixo:

- **Passo 1**: o professor cadastra os alunos, insere o tema, visualiza a discussão dos alunos e também pode ver graficamente os resultados da análise da conversa;
- **Passo 2**: é possível visualizar o contexto da conversa por meio de gráficos;
- **Passo 3**: é possível visualizar os tópicos extraídos da conversa dos alunos;
- **Passo 4**: cada aluno, ao ser cadastrado, recebe um e-mail com os dados de login;
- **Passo 5**: o grupo é a base de dados onde está armazenada a conversa e é acessada pelo AVA;
- **Passo 6**: intrinsecamente o fórum realiza a análise textual.

Figura 20 – Visão geral por meio do modelo de Arquitetura do ambiente desenvolvido.



Fonte: Autoria própria.

A aplicação das técnicas TF e LDA no ambiente virtual diz respeito ao tamanaho do *corpus*. Caso no ambiente os alunos façam um pequeno debate em relação ao tema proposto, a técnica TF poderá suprir a necessidade da análise dos dados textuais, mas se a quantidade de texto for externsa, o LDA se mostra mais eficaz nessa tarefa de análise. Neste ponto é importante lembrar que a técnica PCA já não faz parte das nossas análises porque não foi selecionada como a técnica mais apropriadas para as nossas análises.

4.1.3.2 Aplicação

Para validação do módulo por parte de um docente, foi utilizado o CFS 2020.1 (Curso de Formação de Sargentos 2020.1) da Polícia Militar do Rio Grande do Norte. Esse curso transcorre durante o período de 15 de Junho à 20 de Agosto de 2020. O núcleo de ensino é o Centro de Formação e Aperfeiçoamento da PM (Polícia Militar). A modalidade à distância é uma novidade para esse tipo de curso, ou seja, é a primeira vez que esse curso ocorre EaD na PM RN. Para a validação deste trabalho de dissertação utilizando o CFS, num primeiro contato houve algumas dificuldades para se conseguir uma turma devido às adaptações à nova modalidade. Num segundo contato, foi possível conseguir uma turma para realizar os testes de validação por parte de um docente.

O docente da disciplina Didática Aplicada à Atividade Policial, do CFS 2020.1 da Polícia Militar do Rio Grande do Norte, propôs, durante o período de 10 de julho até 15 de julho, aos alunos sargentos a discussão sobre o título **Ainda somos "cabeça de papel"?**, com o seguinte tema com três perguntas:

- O título do fórum se refere à cantiga popular "Marcha soldado"NOTA: Sem emitir opiniões sobre a fonte do vídeo, observe os questionamentos que surgem, a partir dele:
  - a) Inspirado no texto usado ontem, podemos afirmar que a educação policial militar tem se modificado muito ao longo do tempo. Será que o vídeo proposto condiz ainda com a realidade das ruas na prestação do serviço policial? Explique.
  - b) A solução para prestar um bom serviço para a sociedade potiguar ainda está na educação policial militar? Comente.
  - c) Qual sua ideia sobre o termo "policial-educador"? Refletir sobre a profissão é um exercício sempre necessário.

A partir da análise da discussão proposta, em que se obteveram 697 respostas dos alunos, o modelo probabilístico de tópicos LDA fez uma busca na tarefa de explorar os dados na descoberta dos tópicos, os quais apresentam valor semântico e formam *clusters* que acontecem juntos com frequência. Dessa forma, ao se fazer essa análise, consegue-se inferir um assunto (tema) ou se ter os termos mais relevantes em torno do qual ocorre em um conjunto de textos.

Com a proposta de discussão iniciada, conseguiu-se acompanhar o desenvolvimento da discussão e a definição dos resultados reportados pelo LDA. Após iniciada a discussão proposta, nas primeiras quatro horas de início, obtiveram-se os resultados sobre os termos que concorrem para os tópicos da discussão, conforme se pode observar na Figura 21.

Figura 21 – Os termos mais relevantes por tópicos na discussão.

Fonte: https://bit.ly/37lMaiL.

É importante lembrar que a eficiência dos resultados está diretamente relacionado ao tamanho da amostra, ou seja, quando se aumenta o tamanho da amostra, aumenta-se a relevância estatística que essa amostra pode ter, em outras palavras, estatisticamente é bem menor a chance de cair numa mera coincidência. É valido lembrar também que quanto maior a população dos termos amostrados, melhores são os resultados reportados pelo LDA. Assim, num primeiro momento, a Figura 21 reporta precocemente os termos

que concorrem para os tópicos, porém trazendo significados não muito relevantes para representar a semântica final da discussão. Isso pode ser comparado com a Figura 22, obtida no mesmo dia, porém bem mais tarde.

Figura 22 – Os termos mais relevantes por tópicos na discussão.

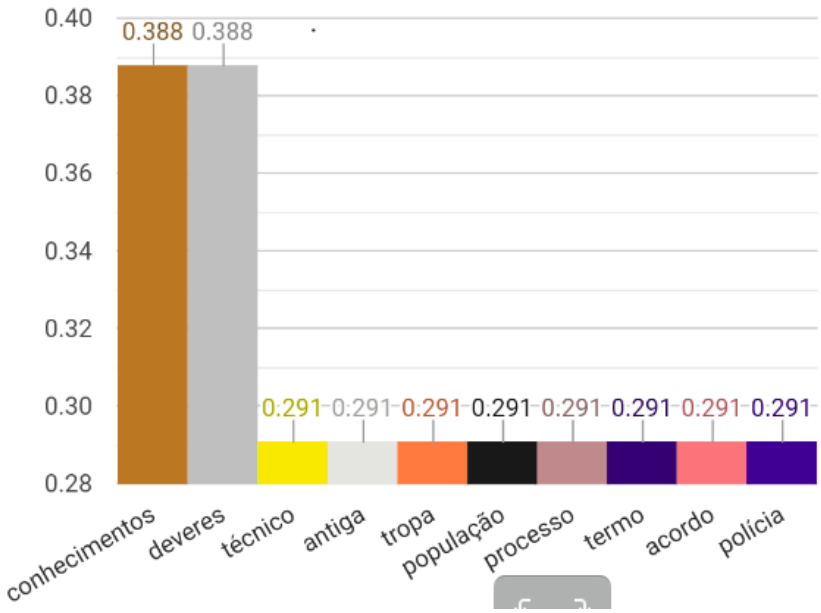
Fonte: https://bit.ly/37lMaiL.

Ao ler o tema proposto pelo docente, percebe-se que a temática que os alunos devem discorrer deve estar em torno dos termos **prestação do serviço policial**, **educação policial militar** e **"policial-educador"**. Com base nesses termos e os comparando aos dois tópicos 1, (Figura 21 e Figura 22), nota-se que o Tópico 1 da Figura 22 mostra relevâcia por parte do tópicos da discussão proposta, haja vista trazer consigo termos associados ao policial militar como **técnico**, **policial-educador** e **educado**. Entre esses dois Tópico 1, os termos que podemos comparar são os termos comum que são *técnico* e *conhecimentos*, como se pode observar nas Figuras Figura 23 e Figura 24. Percebe-se que o aprendizado não supervisiondo do algoritmo LDA vai, num primeiro momento, com a discussão proposta em andamento, coloca os termos **conhecimento** e **técnicos** como sendo mais relevantes.

Continuando a análise das imagens, o tópico 4 da Figura 21 e tópico 2 da Figura 22, pode-se observar que o Tópico 4 mostra algumas palavras que frequentemente ocorrem juntas, são elas: **policial**, **educação**, **serviço**, **formação**, **profissional**, **segurança**, **pública** e **policialeducador**. Após vários outros alunos participarem da conversação, o algoritmo LDA colocar alguns termos que estavam no Tópico 4 e trazendo-os para o Tópico 2, com os termos **policial**, **educação**, **serviço**, **formação**, **segurança**, **militar**, **profissional** e **pública**. A posição do tópico não indica sua relevância, apenas indica que é um *cluster* de termos. Além da comparação dos termos de acordo com o tópico, também se pode observar um aprendizado dentro do próprio tópico de acordo com a conversação. Assim, alguns termos podem mudar sua relevância dentro do próprio tópico, como acontece com o termo **profissional** que perdeu relevância para **segurança**, e **segurança** que perdeu relevância para **militar**.

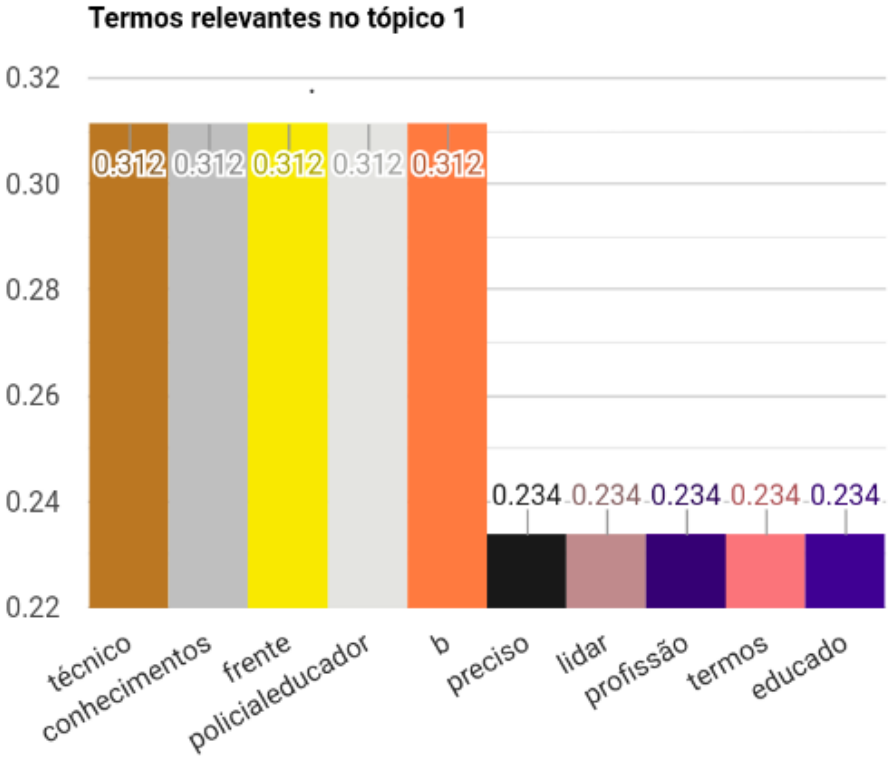
A Figura 25 mostra um exemplo de correlação semântica de termos que aparecem algumas vezes juntos na conversação dos alunos. As palavras **força**, **educacional**, **mediaçao**, **promoção**, **requalificação**, **tiro** e **métodos** parecem não ter valor útil isoladamente,

Figura 23 – Primeira análise do tópico 1 da Figura 21



Fonte: https://bit.ly/37lMaiL.

Figura 24 – Segunda análise do tópico 1 da Figura 22.



Fonte: https://bit.ly/37lMaiL.

mas quando analisadas em relação ao contexto, são carregadas de significados, pois é possível indentificar uma temática como: a **mediação** de **métodos** de **requalificação** impede o uso de **tiro** e até da **força** buscando a **promoção educacional**.

Figura 25 – Exemplo de cluster formado durante a conversação.

força educacional última saem mediação promoção requalificação conteúdos tiro métodos

Fonte: https://bit.ly/37lMaiL.

Outra análise que se pode observar com a aplicação do algoritmo à conversação em fórum, mostra que o modelo LDA faz o reconhecimento de padrões existentes em textos e isso implica que com o amadurecimento da discussão, alguns tópicos ficam quase fixos, como pode ser visto nas figuras Figura 26 e Figura 27.

Figura 26 – Cluster formado no dia 12 de Julho.

Tópico 1: sociedade educação serviço formação vídeo realidade c b educador segurança

Fonte: https://bit.ly/37lMaiL.

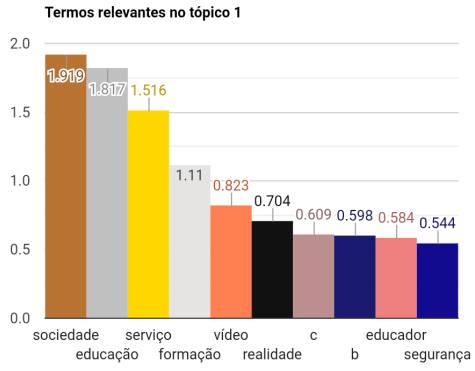
Figura 27 – Cluster formado no dia 13 de Julho.

Tópico 2: sociedade educação serviço formação realidade c b educador segurança prestação

Fonte: https://bit.ly/37lMaiL.

Ainda sobre os tópicos das figuras Figura 26 e Figura 27, pode-se observar também que pouca alteração se obteve sobre a ocorrência de alguns termos. Com isso, é possível observar também, por meio dos gráficos das figuras Figura 28 e Figura 29. Nota-se que alguns termos já pertencem ao tópico e possui aumento nos seus percentuais de relevância com o amadurecimento da discussão entre os alunos.

Figura 28 – Percentuais dos termos da Figura. 26

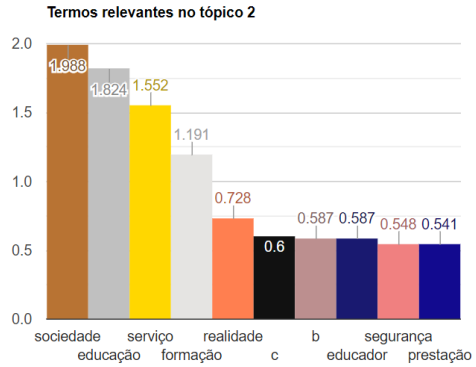


Fonte: https://bit.ly/37lMaiL.

Para melhor entender esse comportamento, a Tabela 23 mostra os termos comuns entre os dois tópicos da Figura 26 e Figura 27, gerdos em momentos diferentes, com seus respectivos percentuais para serem comparados.

Ao longo da conversa dos alunos, como se pode ver na Tabela 23, e estando os mesmos focados no tópico, esses termos tendem a ter uma probabilidade maior de

Figura 29 – Percentuais dos termos da Figura. 27



Fonte: https://bit.ly/37lMaiL.

Tabela 23 – Termos e evolução da aprendizagem do modelo LDA.

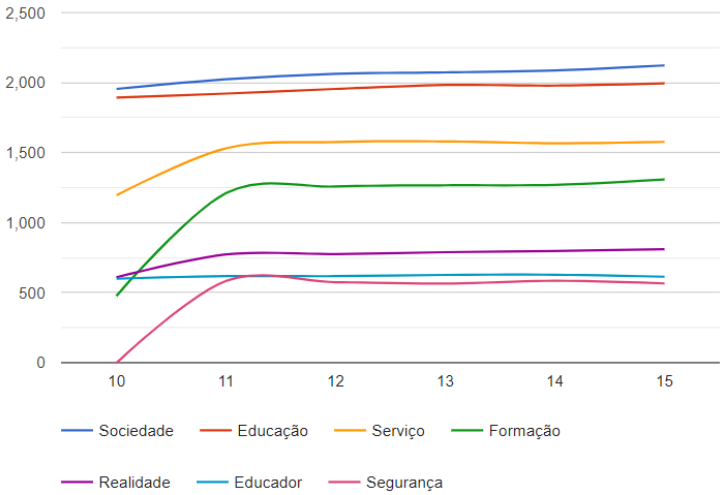
	Sociedade	Educação	Serviço	Formação	Realidade	Educador	Segurança
Tópico 1	1.919	1.817	1.516	1.110	0.704	0.584	0.544
Tópico 2	1.988	1.824	1.552	1.191	0.728	0.587	0.548

Fonte: Autoria própria.

pertencerem ao tópico. Note que sempre um termo no tópico 1 é menor que a sua ocorrência no tópico 2.

Além da Tabela 23 que mostra, à proporção que a discussão transcorre, os termos ganham importância, é possível ver a progressão desses termos, em todos os dias da discussão no fórum, por meio do *Line Chart* (Gráfio de linhas), como se pode ver na Figura 30.

Figura 30 – Progressão da relevância dos termos.



Fonte: https://bit.ly/37lMaiL.

O gráfico da Figura 30 foi realizado por meio de um acompanhamento dos tópicos gerados a cada dia, entre os dias 10 e 15 de julho de 2020. Nesses dias houve bastante postagens dos alunos, como pode ser visto na Tabela 24.

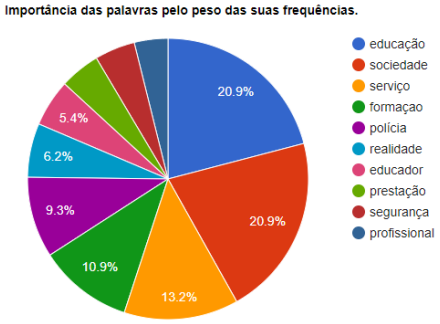
Tabela 24 – Quantidade de postagens por dia de fórum aberto aos alunos.

Dias	10	11	12	13	14	15	**Total**
Postagens	165	157	68	72	103	132	697

Fonte: Autoria própria.

A matriz de frequência gerada no dia 14 de Julho (um dia anterior ao término do fórum) representa os termos com maior peso para aquele dia, como se pode observar no gráfico *Pie Chart* (Gráfico de Pizza) da Figura 31. A ideia por traz da técnica TF não representa palavras que possuam valor semântico, pois cada termo é representado por um peso e esse peso é obtido normalizando a frequência do termo por todo documento.

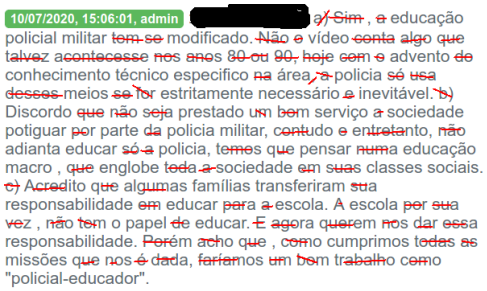
Figura 31 – Importância das 10 palavras pelo peso das suas frequências.



Fonte: https://bit.ly/37lMaiL.

Como visto na RSL 3.1, há possibilidade que haja um ambiente de conversação com uma grande quantidade de termos usados com gíria ou abreviações e quando isso acontece, quando não é possível a exclusão desses termos por *stop words*, o modelo LDA precisa de uma modificação interna. O fórum proposto possibilitou a observação de que as postagens seguiram o padrão normativo de nosso idioma e as abreviações provenientes do jargão militar foram acrescentadas a lista de *stop words*. Na Figura 32 é apresentada uma postagem de um aluno, em que os *stop words* se encontram taxados de vermelho. Outra observação a ser realizada é que, como vimos na RSL 3.1, o modelo LDA tem melhores resultados para grandes quantidades de dados e as 697 postagens possibilitaram explorar os dados de maneira mais eficiente possível.

Figura 32 – Postagem de um aluno com os *stop words* marcados em vermelho.


10/07/2020, 15:06:01, admin [redacted] a) Sim , a educação policial militar tem se modificado. Não o vídeo conta algo que talvez acontecesse nos anos 80 ou 90, hoje com o advento do conhecimento técnico especifico na área, a policia só usa desses meios se for estritamente necessário e inevitável. b) Discordo que não seja prestado um bom serviço a sociedade potiguar por parte da policia militar, contudo e entretanto, não adianta educar só a policia, temos que pensar numa educação macro , que englobe toda a sociedade em suas classes sociais. c) Acredito que algumas famílias transferiram sua responsabilidade em educar para a escola. A escola por sua vez , não tem o papel de educar. E agora querem nos dar essa responsabilidade. Porém acho que , como cumprimos todas as missões que nos é dada, fariamos um bom trabalho como "policial-educador".

Fonte: https://bit.ly/37lMaiL.

Ao analisar todo o desenvolvimento da conversação no fórum, observou-se que a lista de *stop words* inicial para se executar o algoritmo proposto é genérica, haja vista que foi observado que para a área da polícia militar e para um curso de formação CFS, existem muitos termos que são inerentes a essa área de trabalho. Assim, foram adicionados a essa lista termos como **sargento**, **al**, **sgt**, **alsgt**, **aluno**, **mat**, **matrícula**, **QSL**, entre outros termos. Nem todos os verbos são *stop words*, como exemplo tem-se o termo **coagir**, pois nem todo policial pratica a coação sobre alguém, podendo esse termo ser ou não relevante. A retirada e a colocação de um termo como *stop words* pode parecer simples, mas, caso não haja uma boa avaliação da importância de sua ocorrência, pode interferir nos resultados e gerar um erro na conclusão da análise.

Dependendo do contexto, torna-se possível deixar algum *stop word* de maneira que não interfira no resultado final, ou seja, levando a um erro na hora de inferir, ou mesmo os *stop words* deixados implique uma execução exponencial do algoritmo. Deixar alguns termos desses pode ser concedido, pois, às vezes, ajuda o leitor (professor) a entender o significado do contexto de uma conversação. Por isso, para explicar os resultados, este trabalho teve como foco os termos triviais para mostrar a relação do algoritmo LDA com o texto, e esses termos fazem parte do tema, logo são *stop words*. Nesse ponto entra o especialista, a pessoa que sem saber o que é um *stop word*, naturalmente sabe selecionar os termos os quais julgam mais relevantes.

4.1.3.3 Aceitação do Módulo por Parte dos Professores

Esta seção responde à questão de pesquisa *Quais requisitos de aceitação e utilidade percebida do docente?* A pergunta foi realizada após a experiência do professor ter ao módulo de mineração de texto. A fim de se obter a resposta a essa questão de pesquisa, foi realizado o segunte questionário para que o professor falasse sobre essa aceitaçao:

1. Nome completo
    - Gleydson Rodrigues Dantas
2. Cada tópico apresenta palavras que foram escolhidas pela sua relação semântica no contexto da conversação. Qual tópico melhor representa o tema proposto? Quais são as palavras desse tópico?
    - Tópico 1: Conhecimento - Educação - Argumentos - Ignorância - Ciência
3. Cada tópico também apresenta palavras que possuem correlação entre si (que normalmente aparecem juntas) em um texto específico de uma conversação. As palavras que aparecem juntas em um único tópico podem aparecer juntas numa postagem de um aluno? Ou seja, é possível dizer que as palavras de um tópico estão próximas em relação à semântica esperada numa postagem?

- Sim, haja vista tratar-se de temas correlacionados. Sim, estão próximas semanticamente.

4. Antes de iniciar o fórum, se o senhor fosse pensar em algumas palavras que teriam que aparecer nos textos, essas palavras estão presentes nos tópicos mostrados? É possível vê-las no gráfico de pizza?

   - Sim. É possível ver no gráfico.

5. Digamos que o senhor ainda não viu o que os alunos postaram, seria possível, por meio das palavras geradas na parte de estatística inferir que os alunos falavam realmente sobre o tema proposto?

   - Corretamente

6. É possível a aplicação de modelos estatísticos e probabilísticos como esses em um AVA (Ambiente Virtual de Aprendizagem) como o Moodle para ajudar docentes na tarefa de investigar se os alunos estão atentos ao tema proposto para discussão?

   - Desconheço se isso é possível no moodle, mas se houver link direto, sim, seria possível.

7. Com a sua experiência em AVA (Ambiente Virtual de Aprendizagem), é possível a inserção de um módulo estatístico em ambientes como o Moodle que auxilie o professor/tutor a acompanhar o desenvolvimento dos alunos nas discussões do fórum?

   - Por não ser da área de TI, desconheço se isso é possível no moodle.

Com base nessas considerações, nota-se que a Modelagem de Tópicos não pode ser encarada como uma ciência exata como a mátemática. Apenas um especialista é capaz de afirmar se os resultados apresentados são úteis ou totalmente fora de contexto, ou seja, o especialista sabe o que ele tem que está vendo como *feedback*. A resposta a questão 2, supracitada, evidencia que o especialista, pela sua ótica, encontrou um tópico sem os *stop words* do tema proposto.

4.1.3.4 Como analisar se um texto produzido por um grupo de estudo na ferramenta fórum de um AVA está em conformidade com um determinado objetivo de aprendizagem?

Esta pesquisa adotou como uma das metodologias a *design science*, que possibilita alcançar os resultados por meio de respostas a questões de pesquisa. Inicialmente se desenvolveu uma questão geral de pesquisa (QGP) e depois houve um detalhamento com questões específicas, com questões conceituais, tecnológicas e práticas. Essas questões específicas estão respondidas ao longo desta dissertação. A QGP proposta é **como analisar se um texto produzido por um grupo de estudo na ferramenta fórum de um**

AVA está em conformidade com um determinado objetivo de aprendizagem? A resposta a essa QGP foi alcançada por meio de uma RSL, de resultados preliminares, resultados com dados de uma base de dados EAD e por meio diretamente de um fórum de discussão. Dentro do problema de se descobrir estruturas temáticas em torno das quais os alunos de um fórum de um AVA estão debatendo, foi necessária a realização de uma RSL que trouxesse o estado da arte sobre algumas técnicas estatística para se encontrar o quão relevante elas são para esta pesquisa. Por meio dos resultados preliminares em livros do Projeto *Gutenberg*, foi possível verificar que o modelo LDA trouxe termos relacionado à correlação. Com isso, os termos reportados estão lógico-matematicamente próximos, o que possibilita a criação de *clusters* com os termos que estão próximos semanticamente. Dessa forma, se existe a possibilidade de o algoritmo LDA trazer resultados satisfatórios quando aplicados em livros, sua aplicação em fóruns de discussão pode ser testada também. Assim, testou-se o LDA em fóruns e os resultados foram positivos, pois foi possível se extrair a eficácia e a aceitação de maneira positiva.

4.1.3.5 Dificuldades

Encontrar uma turma no modelo EaD focada ao tópico de fóruns não foi uma tarefa fácil inicialmente. Devido à pandemia do Covid-19, a polícia militar do RN teve que dá continuidade as suas atividades letivas e adotou a metodologia EaD. A adoção dessa metodologia não foi fácil, pois muitas barreiras tiveram que ser quebradas. Dessa forma, a polícia militar conseguiu iniciar sua primeira turma e 100% EaD com três plataformas:

- http://www.cfapm.rn.gov.br;
- https://classroom.google.com/;
- http://www2.pm.rn.gov.br/moodle2.

As plataformas do CFAPM e Moodle, nessa primeira fase, não foram bem implementadas para o curso, tendo o mesmo transcorrido em sua maior parte pelo *Google Classroom*. Diante dessa fase de teste a adaptações, conseguir uma turma para aplicar os testes não foi trivial.

O ambiente em forma de fórum desenvolvido para este trabalho de dissertação foi construído utilizando-se a linguagem de programação *PHP*. A escolha dela se deve ao fato de o AVA *Moodle* ser desenvolvido também em *PHP*. A intenção inicial que se tinha era desenvolver toda estrutura do código de mineração de texto. Depois disso, acoplar ou inserir no ambiente do *Moodle*, porém a parte *Developer* do *Moodle Docs*, mostrou-se meio complexa para se conseguir inserir em um Moodle de uma turma. Assim sendo, para se obter resultados em tempo hábil para a defesa deste mestrado, encontrou-se dificuldade para a tarefa de se inserir um módulo de Análise de Dados textual. Isso implicou o

desenvolvimento de um ambiente virtual em forma de fórum. Não foi desenvolvido um ambiente mais amplo em que se teria todos os recursos de um AVA normal, pois o interesse deste trabalho é apenas direcionado à análise dos textos dos alunos.

O fórum tem resultados de duas técnicas: LDA e TF. O LDA é o modelo escolhido após as pesquisas e testes quando comparado com o PCA. O TF é uma técnica tradicional que avalia a relevância do termo pelo seu peso, que está associado a sua frequência. A técnica PCA não se mostrou interessante para o ambiente, haja vista a conclusão dos testes colocar o LDA como a melhor técnica. Mesmo assim, foi realizada uma pesquisa acerca do modelo PCA para se colocar na página dinâmica e os resultados não foram satisfatórios. Isso não trouxe consequências sérias para o trabalho porque não é a técnica de destaque para tratar os dados dos alunos no fórum.

Mesmo não se conseguindo colocar o modelo LDA em um AVA tradicional, é relevante que a proposta seja validada em outro ambiente virtual além do implementado neste trabalho. Para isso, fica como trabalhos futuros para mim ou para outro aluno de mestrado a aplicação do módulo diretamente no AVA.

5 Considerações Finais

Este trabalho de dissertação, sobre o uso da ciência de dados, propôs conseguir inferir se alunos de um AVA, estão dentro da temática escolhida por um docente, para debate em um fórum de discussão. Por meio desse objetivo, conseguiu-se fazer pesquisas e testes que mostraram a melhor técnica entre a LDA e a PCA para a mineração de textos em um fórum de um AVA. Para alcançar esse objetivo, foi necessária a construção de um fórum de discussão. Esse artefado em forma de módulo não está voltado à metaciência e, dessa forma, possibilita algo que acrescente conhecimento a tecnologias aplicadas à educação. A linha de pesquisa proposta tem como objetivo analisar se os alunos estão dentro de um tema que um docente propõe para discussão em grupo.

Por meio de uma RSL, em 3.1, foi realizado uma busca por métodos, observações práticas de análise, testes e comprovações, a fim de capturar informações sobre os efeitos que as técnicas LDA e PCA tem sobre o fenômeno mineração de texto. A partir disso, obteve-se resultados consistentes que fornecem evidências, de que a utilização delas se apresenta robusta o suficiente para trazer respostas coerentes no universo da redução da dimensão de dados na forma textual. Isso contribuiu para descoberta de conhecimento de modo que se possa comparar as técnicas propostas. As pesquisas realizadas neste trabalho possbilitaram a escolha da melhor técnica para o fenômeno mineração de texto para um fórum de um AVA.

Neste trabalho foi utilizado a estatística por intermédio de um modelo probabilístico de tópicos e de um modelo da estatística multivariada para auxiliar a análise textual com a utilização dos métodos LDA e PCA, respectivamente, a fim de fazer uma comparação entre eles e conseguiu-se mostrar os resultados da análise de desempenho, em termos de acurácia, LDA e PCA reportaram resultados satísfatório, mas o LDA se apresentou como a melhor devido à sua filosofia técnica de trazer os termos que estão semanticamente correlacionados. A importância dessa descoberta é relevante, haja vista existir muitos métodos para análise textual. Este trabalho mostra de fato qual método apresenta acurácia melhor quando o assunto é mineração de texto para estruturas formais. Dessa forma, foi possível mostrar os resultados que a estatística pode oferecer para descoberta de características em textos por meio da utilização e comparação das duas técnicas propostas.

A validação do modelo LDA, por meio de um fórum de aprendizagem, corroborou com a RLS e com dos testes preliminares, pois o especialista não apenas aprovou os resultados, mas ainda destacou que seria possível a implantação desse módulo em um AVA, para ajudar docentes na tarefa de investigar se os alunos estão atentos ao tema proposto para discussão.

Até este momento, este trabalho deixou uma lacuna a respeito da possibilidade de se gerar tópicos a partir das componentes principais, ou seja, pegando os dados na forma de texto já tratados e aplicando a técnica PCA para reduzir a dimensionalidade e depois, em cima das componentes geradas, aplicar-se o método LDA, e verificar se o resultado está dentro do esperado. Dessa forma, pode-se deixar uma questão de pesquisa para trabalhos futuros: **é possível analisar o desempenho de uma abordagem híbrida que combina as técnicas LDA e PCA**?

É importante ressaltar que os resultados **preliminares** e os resultados das **bases de dados AVA** apresentados não possuem o mesmo comportamento para outros tipos de textos. Os resultados apresentados são baseados a partir de estruturas textuais que seguem as gramáticas normativas. Os livros do projeto *Gutemberg* possuem um tamanho considerável para uma obra, enquanto que os dados dos AVA possuem tamanho variável que denpende da interação nas discussões. Ressalta-se também que textos curtos (que são suscetíveis a gírias e abreviações) ou documentos formais, que possuem um estrutura peculiar, podem apresentar resultados diferentes daqueles da seção Resultados Preliminares 2(b)i.

Como não foi conseguido em tempo hábil a inserção do módulo de mineração de texto em um AVA, é possível deixar para trabalhos futuros essa tarefa já testada e aplicada neste trabalho.

6 Resultados Esperados

Com este trabalho, objetiva-se alcançar os seguintes resultados:

1. Mostrar resultados referentes à analise de dados textuais em fóruns de discussão para AVA;
2. Apresentar a melhor técnica estudada neste trabalho de dissertação para um módulo de mineração de texto em fóruns de AVA;
3. Orientar pesquisadores do fenômeno mineração de texto para fóruns de discussão em AVA.
4. Melhorar o desempenho dos estudantes em fóruns de discussão em um ambiente AVA a partir de uma avaliação do que eles estão debantendo se está dentro do contexto proposto.

Referências

AKEN, J. E. V.; ROMME, G. Reinventing the future: adding design science to the repertoire of organization and management studies. *Organization Management Journal*, Taylor & Francis, v. 6, n. 1, p. 5–12, 2009. Citado na página 14.

ALAN, R. H. V. *et al.* Design science in information systems research. *MIS quarterly*, Springer, v. 28, n. 1, p. 75–105, 2004. Citado na página 14.

ALBERTI, T. F. *et al.* Dinâmicas de grupo orientadas pelas atividades de estudo: desenvolvimento de habilidades e competências na educação profissional. *Revista Brasileira de Estudos Pedagógicos*, v. 95, n. 240, 2014. Citado na página 11.

Alhawarat, M.; Hegazi, M. Revisiting k-means and topic modeling, a comparison study to cluster arabic documents. *IEEE Access*, v. 6, p. 42740–42749, 2018. ISSN 2169-3536. Citado na página 48.

Ali, M.; Khalid, S.; Aslam, M. H. Pattern based comprehensive urdu stemmer and short text classification. *IEEE Access*, v. 6, p. 7374–7389, 2018. ISSN 2169-3536. Citado na página 21.

ALMEHMADI, A.; JOUDAKI, Z.; JALALI, R. Language usage on twitter predicts crime rates. In: ACM. *Proceedings of the 10th International Conference on Security of Information and Networks*. [S.l.], 2017. p. 307–310. Citado 2 vezes nas páginas 52 e 53.

BAKHSHAEI, S.; SAFABAKHSH, R.; KHADIVI, S. Extracting parallel fragments from comparable documents using a generative model. *Computer Speech Language*, v. 53, p. 25 – 42, 2019. ISSN 0885-2308. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0885230815300589>. Citado na página 48.

BALIKAS, G.; AMINI, M.-R.; CLAUSEL, M. On a topic model for sentences. In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2016. (SIGIR '16), p. 921–924. ISBN 978-1-4503-4069-4. Disponível em: <http://doi.acm.org/10.1145/2911451.2914714>. Citado 2 vezes nas páginas 48 e 49.

BARTAL, Y.; CHARIKAR, M.; INDYK, P. On page migration and other related task systems. In: CITESEER. *SODA*. [S.l.], 1997. p. 43–52. Citado na página 23.

BIOLCHINI, J. C. de A. *et al.* Scientific research ontology to support systematic review in software engineering. *Advanced Engineering Informatics*, Elsevier, v. 21, n. 2, p. 133–151, 2007. Citado na página 35.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado 4 vezes nas páginas 19, 26, 27 e 30.

CASELLA, G.; FIENBERG, S.; OLKIN, I. Springer texts in statistics. Springer, 2013. Citado na página 66.

CERQUITELLI, T. *et al.* Data miners' little helper: Data transformation activity cues for cluster analysis on document collections. In: *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*. New York, NY, USA: ACM, 2017. (WIMS '17), p. 27:1–27:6. ISBN 978-1-4503-5225-3. Disponível em: <http://doi.acm.org/10.1145/3102254.3102288>. Citado 3 vezes nas páginas 48, 49 e 50.

CHENG, C.-H.; HUNG, W.-L. Tea in benefits of health: A literature analysis using text mining and latent dirichlet allocation. In: *Proceedings of the 2Nd International Conference on Medical and Health Informatics*. New York, NY, USA: ACM, 2018. (ICMHI '18), p. 148–155. ISBN 978-1-4503-6389-1. Disponível em: <http://doi.acm.org/10.1145/3239438.3239459>. Citado na página 49.

DARÓCZI, G. *Mastering data analysis with R*. [S.l.]: Packt Publishing Ltd, 2015. Citado na página 59.

FALEIROS, T. d. P.; LOPES, A. d. A. *et al.* Modelos probabilísticos de tópicos: desvendando o latent dirichlet allocation. São Carlos, SP, Brasil., 2016. Citado 3 vezes nas páginas 23, 27 e 30.

FELDMAN, R.; SANGER, J. *The text mining handbook: advanced approaches in analyzing unstructured data*. [S.l.]: Cambridge university press, 2007. Citado 2 vezes nas páginas 21 e 34.

FLISAR, J.; PODGORELEC, V. Document enrichment using dbpedia ontology for short text classification. In: ACM. *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. [S.l.], 2018. p. 8. Citado na página 53.

JAMES, G. *et al. An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. Citado na página 23.

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, v. 33, n. 2004, p. 1–26, 2004. Citado 4 vezes nas páginas 34, 35, 39 e 40.

KWARTLER, T. *Text mining in practice with R*. [S.l.]: John Wiley & Sons, 2017. Citado na página 57.

Lin, C. *et al.* Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, v. 24, n. 6, p. 1134–1145, June 2012. ISSN 1041-4347. Citado na página 51.

Martinez, A. M.; Kak, A. C. Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 23, n. 2, p. 228–233, Feb 2001. ISSN 0162-8828. Citado 3 vezes nas páginas 47, 48 e 49.

MILLER, J. D. *Statistics for Data Science: Leverage the power of statistics for Data Analysis, Classification, Regression, Machine Learning, and Neural Networks*. [S.l.]: Packt Publishing Ltd, 2017. Citado 2 vezes nas páginas 19 e 30.

MIRIAM. *Very basic strategies for interpreting results from the Topic Modeling Tool*. 2012. Acessado: 26 de Fevereiro de 2020. Disponível em: <http://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/>. Citado na página 24.

MONTGOMERY, D. C.; RUNGER, G. C. *Estatística Aplicada e Probabilidade para Engenheiros. 5ª*. [S.l.: s.n.], 2012. Citado 3 vezes nas páginas 22, 23 e 25.

MUNZERT, S. *et al. Automated data collection with R: A practical guide to web scraping and text mining*. [S.l.]: John Wiley & Sons, 2014. Citado 2 vezes nas páginas 19 e 21.

NETO, J. M. M. Estatística multivariada-uma visão didática-metodológica. *Crítica Revista de Filosofia, maio*, 2004. Citado na página 30.

OLIVEIRA, P. F. d.; GUERRA, S.; MCDONNELL, R. *Ciência de Dados com R. 2ª*. [S.l.: s.n.], 2018. Citado 4 vezes nas páginas 16, 17, 34 e 55.

PARK, K.; NGUYEN, M. C.; WON, H. Web-based collaborative big data analytics on big data as a service platform. In: IEEE. *2015 17th International Conference on Advanced Communication Technology (ICACT)*. [S.l.], 2015. p. 564–567. Citado na página 33.

RODRIGUES, A.; COELHO, A. C.; PAULO, E. Análise multivariada para os cursos de administração, ciências contábeis e economia. 2007. Citado 2 vezes nas páginas 24 e 30.

SANCHETI, P.; SHEDGE, R.; PULGAM, N. Word-ipca: An improvement in dimension reduction techniques. In: IEEE. *2018 International Conference on Control, Power, Communication and Computing Technologies (ICCPCCT)*. [S.l.], 2018. p. 575–578. Citado na página 20.

SOUZA, T. I. A. de. *Aplicação de Técnicas Supervisionadas e Não-Supervisionadas da Estatística Multivariada no Contexto da Avaliação da Efetividade do Ensino em um Curso de Graduação*. 129 p. Dissertação (Mestrado) — Universidade Federal do Ceará, http://www.repositorio.ufc.br/handle/riufc/22972, 2016. Citado na página 12.

STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. *Handbook of latent semantic analysis*, v. 427, n. 7, p. 424–440, 2007. Citado 2 vezes nas páginas 24 e 25.

Sushma Niket Borade; Adgaonkar, R. P. Comparative analysis of pca and lda. In: *2011 International Conference on Business, Engineering and Industrial Applications*. [S.l.: s.n.], 2011. p. 203–206. Citado 2 vezes nas páginas 47 e 49.

WICKHAM, H.; GROLEMUND, G. *R for data science: import, tidy, transform, visualize, and model data*. [S.l.]: "O'Reilly Media, Inc.", 2016. Citado 3 vezes nas páginas 16, 18 e 55.

WIERINGA, R. J. What is design science? In: *Design Science Methodology for Information Systems and Software Engineering*. [S.l.]: Springer, 2014. p. 3–11. Citado na página 54.

Zare, A. *et al.* Extension of pca to higher order data structures: An introduction to tensors, tensor decompositions, and tensor pca. *Proceedings of the IEEE*, v. 106, n. 8, p. 1341–1358, Aug 2018. ISSN 0018-9219. Citado 2 vezes nas páginas 23 e 30.

ZENG, J.; CHEUNG, W. K.; LIU, J. Learning topic models by belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 35, n. 5, p. 1121–1134, 2012. Citado 2 vezes nas páginas 33 e 48.

ZHANG, Y.; MAO, W.; ZENG, D. A non-parametric topic model for short texts incorporating word coherence knowledge. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2016. (CIKM '16), p. 2017–2020. ISBN 978-1-4503-4073-1. Disponível em: <http://doi.acm.org/10.1145/2983323.2983898>. Citado 3 vezes nas páginas 48, 50 e 52.