UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MÉTODO SUPERVISIONADO PARA IDENTIFICAÇÃO DE DÚVIDAS EM POSTAGENS DE FÓRUNS EDUCACIONAIS

VITOR BELARMINO ROLIM

RECIFE

2016

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

VITOR BELARMINO ROLIM

MÉTODO SUPERVISIONADO PARA IDENTIFICAÇÃO DE DÚVIDAS EM POSTAGENS DE FÓRUNS EDUCACIONAIS

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

Orientador: Prof. Dr. Rafael Ferreira Leite de Mello
Co-orientador: Prof. Dr. Evandro de Barros Costa

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

FICHA DE APROVAÇÃO DO TRABALHO DE CONCLUSÃO DE CURSO

Trabalho defendido por Vítor Belarmino Rolim como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado *Método Supervisionado para identificação de dúvidas em postagens de fóruns educacionais*, orientado por Rafael Ferreira Leite de Mello e aprovado pela seguinte banca examinadora:

Rafael Ferreira Leite de Mello
DEINFO/UFRPE

André Câmara Alves do Nascimento
DEINFO/UFRPE

Filipe Rolim Cordeiro
DEINFO/UFRPE

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter me dado, acima de tudo, paciência para suportar esta jornada.

Aos meus pais, Flávio e Ilma, por todo amor e apoio durante toda a minha vida, incondicionalmente. E por fazer sempre o que estivesse ao alcance para proporcionar uma boa educação.

A minha irmã Flávia e ao meu cunhado Júnior, por sempre estarem presentes, e por todos os momentos de descontração.

Ao meu orientador, Rafael Ferreira, por ter me impulsionado no mundo acadêmico, e por toda orientação e suporte sempre que necessário.

Ao corpo docente do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco, por terem colaborado na minha formação acadêmica.

Aos meus colegas de turma, pelo companheirismo.

RESUMO

Com o crescimento da educação à distância, e o uso da tecnologia como ferramenta de apoio educacional, a utilização dos ambientes virtuais de aprendizagem também cresceu. Esses ambientes oferecem vários recursos que podem ser utilizados para auxiliar no processo ensino-aprendizagem, um desses recursos é o fórum. O uso de fóruns de discussão em plataformas educacionais online requer muitas vezes a necessidade de acompanhamento de milhares de usuários. Devido à grande quantidade de postagens, é difícil para o professor e tutor acompanhar os alunos. Para esse acompanhamento ser eficiente, é necessário dispor de ferramentas que auxiliem o professor. Este trabalho tem como objetivo desenvolver uma abordagem para o gerenciamento das postagens dos fóruns educacionais para auxiliar o professor. Para atingir esse objetivo foram realizadas duas etapas: identificação de postagens de dúvida e extração do assunto da postagem. Na etapa da identificação o algoritmo que mostrou um melhor percentual de acerto foi a rede neural (*MultilayerPerceptron),* atingindo um percentual de acerto na classificação de 97%. Na etapa da extração dos assuntos, foi implementado um algoritmo para extração dos assuntos das postagens classificadas como dúvidas. Este algoritmo alcançou uma taxa de acerto de 76,1%. Além disso, o sistema recomenda um vídeo para o aluno de acordo com o assunto extraído da postagem. Essa abordagem auxilia tanto o professor, reduzindo o tempo empregado para responder todos os questionamentos dos alunos, quanto o aluno, fornecendo materiais de estudo para auxiliar a resolver o seu questionamento.

Palavras-chave: Fóruns educacionais, classificação de texto, extração de assunto, recomendação.

ABSTRACT

With the growth of distance education, and the adoption of computational systems as an educational support tool, the use of virtual learning environments has also grown. These environments offer various tools that could be used to assist in the teaching-learning process, one of these tools is the forum. Discussion forums in online educational platforms often requires monitoring of thousands of users. Due to the large number of posts, it is hard to the teacher and tutor supervise the students. In order to perform an effective monitoring, it is necessary tools to assist the teacher. This study aims to develop an approach to manage educational forums posts. To achieve this goal, two steps were performed: the identification of posts as doubt, and the extraction of the subject from post. In the Identification stage, the neural network (MultilayerPerceptron) algorithm achieve better results reaching a 97%. In subject extraction step, it was implemented an algorithm for extracting subjects of posts classified as doubts. This algorithm achieved a 76.1% of accuracy. Moreover, the system recommended videos to the student according to the subject extracted from posts. This approach assists the teacher by reducing the time used to answer the students' questions, and the student by providing supplementary materials to help solve his doubt.

Keywords: Educational forums, text classification, subject extraction, recommendation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

AVA	Ambiente Virtual de Aprendizagem
UFAL	Universidade Federal de Alagoas
UFRPE	Universidade Federal Rural de Pernambuco
CRISP-DM	*Cross Industry Standard Process for Data Mining*
MLP	*Multilayer Perceptron*
TF	*Term Frequency*
IDF	*Inverse Document Frequency*
SVM	*Support Vector Machine*
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1. INTRODUÇÃO

Com o crescente uso da tecnologia como ferramenta de apoio educacional, o uso de Ambientes Virtuais de Aprendizagem (AVA) [Dillenbourg e Schneider 2002] tem aumentado nos últimos anos. Estes ambientes disponibilizam várias ferramentas para melhorar a interação entre professores e alunos, onde alguns exemplos são: fórum, blog, *wiki*, redes sociais, entre outros.

Estas ferramentas possuem um grande potencial para gerar conteúdo, o que pode ser usado para auxiliar no processo ensino-aprendizagem. Porém, devido à grande quantidade de interações entre os alunos e o professor, torna-se difícil para o professor proporcionar um auxilio personalizado para um determinado aluno, por isso é importante que os AVAs ofereçam meios de acompanhamento direto e indireto para garantir o aprendizado do aluno [Akyuz and Kurt 2010].

O acompanhamento direto é aquele realizado sob a supervisão do professor ou tutor. Para isso, o curso tem um plano de ensino e cronograma de atividades que são acompanhados de perto pelos professores e tutores. Então todo material e discussões disponibilizados no AVA é verificado manualmente para que as dúvidas dos alunos sejam resolvidas e seus progressos computados. Devido à grande quantidade de informação, esse tipo de acompanhamento se torna por vezes inviável de ser realizado.

Para amenizar essa situação, é necessário também realizar o acompanhamento indireto, que é o acompanhamento sem a participação direta do professor. Para isso, é importante ter sistemas automatizados que possam auxiliar o professor nessa atividade.

A ferramenta assíncrona de fórum tem uma característica importante, é nela que os alunos postam dúvidas, comentários sobre a disciplina, outras fontes de assunto, possíveis respostas para questões levantadas pelo professor, entre outros. Cheng *et al*. [Cheng *et al.* 2011] realizou um estudo que mostra a efetividade da ferramenta de fórum para melhorar a performance de estudantes em um AVA.

Em um acompanhamento direto as postagens de dúvidas devem receber uma solução. Por outro lado, uma resposta pode ser usada para perceber o progresso do

aluno, com isso o professor pode pontuá-lo ou pode utilizar o aluno como propagador do assunto entre os colegas [Kim 2013].

Para realizar o acompanhamento indireto de um fórum educacional de forma eficiente é importante a utilização de sistemas automáticos [Rolim *et al.* 2014, Azevedo *et al.* 2011, Gerosa *et. al.* 2003]. Por exemplo, um sistema para recomendar automaticamente materiais direcionados para as dúvidas de cada aluno [Baker e Yacef 2009, Mohamad e Tasir 2013], ou um sistema que indicasse ao professor as postagens mais relevantes de acordo com o assunto abordado.

Este trabalho terá como principal objetivo propor um sistema para acompanhamento indireto de fóruns educacionais. Para isso, o primeiro passo é realizar a identificação do tipo de postagem feita no fórum educacional. O segundo passo é extrair o assunto contido na postagem. Uma vez identificado o tipo de postagem e o assunto que ela possui, se a postagem for uma dúvida, direciona conteúdos relacionados ao assunto extraído. Este sistema identifica automaticamente postagens que correspondem a uma dúvida, extrai o assunto abordado por uma determinada postagem e recomenda um material de estudo que possa ajudar ao aluno.

1.1 JUSTIFICATIVA

O fórum é uma ferramenta de comunicação onde vários usuários interagem de forma assíncrona. Desta forma é possível que o usuário faça a sua intervenção de forma mais organizada. No contexto educacional, o fórum abre várias possibilidades para professores e tutores interagirem de forma efetiva com a turma.

Um aspecto importante é que o fórum é considerado uma das ferramentas que permite maior interatividade entre alunos e professores. Segundo o estudo promovido por Barros e Carvalho [Barros e Carvalho 2011] o fórum foi apontado por 69,2% dos alunos como a ferramenta mais interativa, outras ferramentas que também tem essa característica são: tarefa (41,0%), *chat* (38,5%) e questionário (20,5%).

O professor pode usar o fórum em diversos contextos, por exemplo [Freitas 2009]:

(i) incentivar a criação de laços entres os alunos a partir da discursão de temas específicos da disciplina;
(ii) desenvolver a capacidade de debate crítico acerca de algum tema ou assunto;
(iii) dar uma resposta sobre dúvidas e comentários;
(iv) guiar os estudos dos alunos baseados nas suas postagens;
(v) avaliar o aluno.

O ganho pedagógico causado pela utilização dos fóruns educacionais é perceptível, mesmo que os alunos envolvidos no processo empreguem pouco tempo para utilizar esta ferramenta [Cheng *et al.* 2011].

Contudo, com o aumento da interação entre os usuários do fórum, ocorre um aumento também do volume das postagens, tornando-se por vezes inviável o gerenciamento de todas elas pelo professor, demandando muito tempo do mesmo. Por isso faz-se necessária a existência de uma ferramenta que auxilie a análise das postagens dos fóruns educacionais, como o objetivo de maximizar o ganho de conhecimento [Azevedo *et al.* 2011].

Tendo conhecimento de todas essas informações, identificou-se a necessidade da implementação de uma ferramenta que possa auxiliar o professor no processo de aprendizado dos alunos nos AVAs.

1.2 OBJETIVOS

Esta seção contém os objetivos gerais e específicos deste trabalho.

1.2.1 Geral

Este trabalho tem como objetivo, desenvolver uma abordagem para o gerenciamento das postagens dos fóruns educacionais, auxiliando o professor no acompanhamento indireto, que fará a identificação de dúvidas, a identificação do conteúdo da mesma, e a recomendação de um material de estudo auxiliar. Esta abordagem proporcionará o uma melhora no acompanhamento indireto, por recomendar conteúdos para as dúvidas dos alunos automaticamente.

1.2.2 Específicos

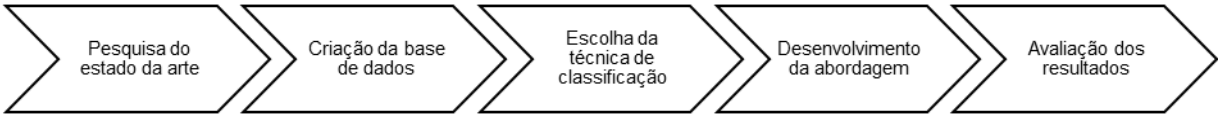
Dos objetivos específicos deste trabalho, podemos destacar os seguintes pontos:

- Construção de um banco de dados de postagens de fóruns educacionais anotadas em dúvidas ou repostas.
- Analisar e comparar as técnicas de classificação: Árvores de decisão, Naive Bayes e Redes neurais, aplicadas na identificação de dúvidas em fórum.
- Classificar as postagens da base de dados utilizando a técnica melhor avaliada.
- Extrair o conteúdo das postagens de dúvida para identificar o tema abordado, a fim de recomendar conteúdo complementar que ajude o aluno a lidar com uma dúvida.

1.3 DEFINIÇÃO DA METODOLOGIA

Podemos observar as etapas do desenvolvimento deste trabalho na figura 1 logo abaixo.

Figura 1 - Esquema de realização do trabalho.



Fonte: O autor.

Primeiramente foi realizada uma extensa pesquisa por trabalhos relacionados a este trabalho, com a finalidade de embasar e justificar a escolha do tema deste trabalho.

Em segundo lugar, foi necessário a construção de 3 bases de dados diferentes, totalizando 1487 postagens.

A necessidade de três bases de dados diferentes, se deve à variação de testes que foram realizados para validação dos resultados. Mais detalhes sobre as bases de dados podem ser encontrados no capítulo 5 deste trabalho.

Em seguida, um estudo foi realizado para definir a melhor técnica de classificação que se aplicaria ao problema abordado neste trabalho. Esta fase de estudos e análises da melhor técnica para resolução do problema pode ser encontrada no trabalho de Rolim *et al.* 2014 (vide anexo III).

Posteriormente ao estudo das técnicas de classificação, foi realizado o desenvolvimento da abordagem proposta. Esse desenvolvimento se dividiu em 3 etapas: classificação das postagens, extração dos assuntos, e recomendação de material de estudo. Todas as etapas foram desenvolvidas com a linguagem de programação *Python*. Detalhes do desenvolvimento podem ser encontrados no capítulo 4.

Como última etapa deste trabalho, foi realizada a avaliação dos resultados obtidos.

1.4 ESTRUTURA DO TRABALHO

Este trabalho divide-se da seguinte forma. No primeiro capítulo foi apresentada a introdução, objetivos e justificativas do tema abordado, bem como a metodologia adotada. O segundo capítulo explana sobre a fundamentação teórica deste trabalho. No capítulo três, os trabalhos relacionados são apresentados. O capítulo quatro mostra como foi desenvolvida a proposta desse trabalho. O capítulo cinco apresenta os resultados obtidos dos testes realizados, a fim de avaliar a abordagem desenvolvida. No sexto capítulo as conclusões sobre o trabalho são apresentadas, assim como a proposta de trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda os temas necessários para a melhor compreensão dos tópicos abordados neste trabalho.

2.1 CLASSIFICAÇÃO

Aprendizagem de máquina é uma das áreas da inteligência artificial, podemos definir como sistemas capazes de adquirir conhecimento a partir de dados. Weiss e Kulikowski [Weiss e Kulikowski 1991] definem um sistema de aprendizado de máquina como um programa de computador que toma decisões baseadas na experiência contida em exemplos solucionados com sucesso.

Uma das categorias do aprendizado de máquina é o aprendizado supervisionado, que tem como objetivo compreender a relação entre os parâmetros fornecidos, para poder classificar, ou etiquetar, uma determinada instância (imagem, documento, etc.). Tem diversas aplicações, uma delas é a classificação textual.

A classificação textual tem como objetivo determinar se um documento específico pertence a uma ou mais classes [Sebastiani 2002]. Este processo é chamado de classificação ou categorização.

Classificação textual pode ser considerada uma abordagem que consiste em separar previamente uma coleção de documentos, identificando-os como pertencentes a uma classe específica. Posteriormente, um algoritmo de aprendizagem de máquina ou algoritmo de classificação é treinado com essa coleção como dado de entrada. A partir deste treinamento o classificador é capaz de indicar a classe de uma entrada fornecida como teste.

Podemos citar várias técnicas de aprendizagem de máquina, tais como: Naive Bayes, redes neurais, árvore de decisão, entre outras.

2.1.1 Naive Bayes

Naive Bayes é uma técnica probabilística baseada no Teorema de Bayes [Rutten 2010] e é uma técnica bastante utilizada em aprendizagem de máquina e reconhecimento de padrões. Esse classificador calcula a probabilidade de que uma

amostra desconhecida pertença a cada uma das classes possíveis, predizendo a classe mais provável. Para isto, o classificador bayesiano calcula uma distribuição geradora para cada classe do problema através da análise das relações entre as características envolvidas e as classes de cada instância.

Após treinar um classificador Bayesiano, dado uma amostra ele decide pela classe com maior probabilidade, representado por $P(w_i/x)$. Essa probabilidade é calculada pela equação 1.

$$P(w_i/x) = \frac{P(w_i)\rho(x/w_i)}{\rho(x)}, \quad (1)$$

$$\rho(x) = \sum_{i=1}^{c} P(w_i)\rho(x/w_i), \quad (2)$$

onde $\rho(x)$, representada pela equação 2, é a função de densidade de probabilidade das classes e $\rho(x/w_i)$ é a função de probabilidade de cada classe w_i. O classificador bayesiano realiza uma classificação estatística, sendo completamente baseado em probabilidades. Dentre os classificadores bayesianos, podemos citar o NaiveBayes [McCallum e Nigam 1998]

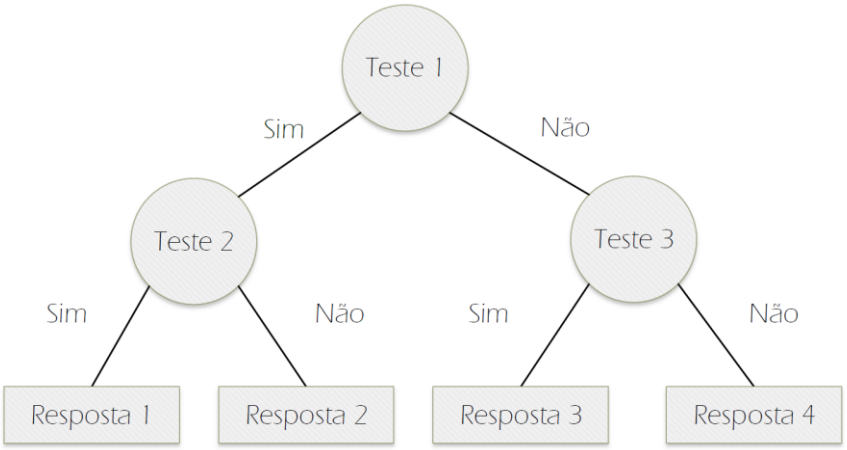
2.1.2 Árvore de decisão

Árvore de decisão [Bhargava *et al.* 2013] é uma técnica de aprendizado de máquina que utiliza uma estrutura de árvore para avaliar os atributos de uma entrada e retorna uma predição baseada nos valores desses atributos. Essa árvore é estruturada através de vários nós, onde cada nó corresponde a um teste do valor de uma característica do dado de entrada. Os nós da árvore são ligados por ramos, os quais identificam os possíveis valores do teste realizado em cada nó. Por fim, cada nó da folha da árvore representa um valor de retorno. Sendo assim, a árvore de decisão chega a uma decisão através da realização de vários testes. Para isso, o algoritmo é iniciado na raíz da árvore e a percorre realizando testes sobre as características, que correspondem aos nós, do dado de entrada até chegar na folha da árvore. Ao chegar na folha da árvore é retornado como resultado a classificação.

A Figura 2 mostra a estrutura de uma árvore de decisão binária. Em cada nó da árvore é realizado um teste baseado nos valores dos atributos da amostra e o

algoritmo percorre a árvore até chegar na folha, onde é retornado uma resposta de classificação. A árvore de decisão utilizada no trabalho, no entanto, não é binária, podendo ter mais de uma saída para cada nó da árvore. Os testes realizados em cada nó, para o trabalho proposto, envolvem os valores das características selecionadas nos dados de entradas. A resposta da árvore, no caso proposto, é a classe da postagem de entrada no classificador.

Figura 2 - Estrutura de uma árvore de decisão.



Fonte: Rolim *et al.* 2014.

Para medir se a habilidade de um dado atributo é o mais adequado para realizar os testes em cada nó e ser usado como classificador utiliza-se duas métricas: a entropia e o ganho de informação. A entropia mede o nível de incerteza (ou impureza) de um conjunto. O ganho de informação (GI) que é o responsável por determinar qual é de fato o melhor atributo.

$$Entropia(S) = -p_{\oplus} log_2 p_{\oplus} - p_{\ominus} log_2 p_{\ominus} \quad (3)$$

$$GI(S, A) = E(S) - \sum_{veValores(A)} \frac{S_v}{S} E(S) \quad (4)$$

Na equação 3 observamos a representação de entropia, onde o *S* é uma amostra dos exemplos de treinamento, p_+ é a proporção de exemplos positivos em *S*, e p_- é a proporção de exemplos negativos em *S.* E na equação 4 observamos a fórmula do GI que é calculada a partir do conhecimento dos valores da entropia do

conjunto *S*. O atributo com o maior valor de GI é selecionado, pois é o que mais reduz o nível de incerteza.

Existem algumas variações de implementação de árvores de decisão, uma delas é a árvore de decisão J48 [Bhargava *et al.* 2013].

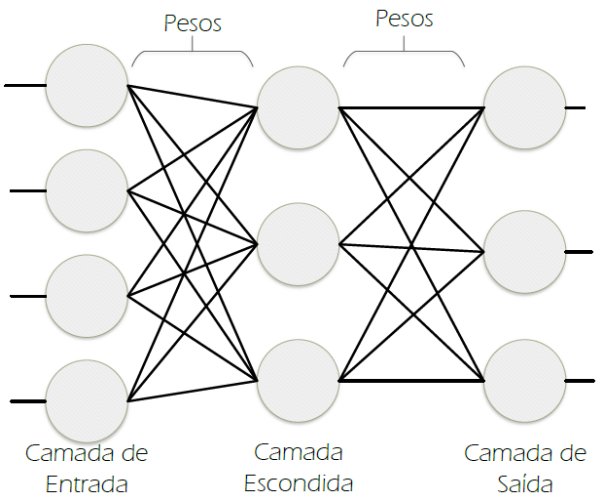
2.1.3 Rede neural

A rede neural [Haykin 1999] é um modelo computacional inspirado nas ligações entre neurônios do cérebro humano. Esse modelo é composto por um conjunto de neurônios, ou nós, que são interligados entre si, formando uma rede. Esses neurônios são divididos em camadas e são conectados por meio de ligações. Cada ligação possui um peso (fórmula para cálculo do peso na equação 5), os nós são ajustados baseados nas características dos dados de entrada. A Figura 3 mostra a estrutura de uma rede neural.

$$w_i = w_i - \alpha \delta_j^i (y_j^{i-1})^T \quad (5)$$

A partir dos dados de entrada e da força de ligação entre os nós, a rede realiza a classificação dos dados entrada. No trabalho proposto são utilizados 4 neurônios na camada de entrada, onde para cada nó é atribuído uma característica da postagem a ser classificada. Como são extraídas quatro características para cada postagem, que serão descritas no capítulo 4, a camada de entrada é composta por 4 neurônios. A camada de saída apresenta três neurônios, onde cada neurônio corresponde a uma classe: pergunta, dúvida ou resposta. O neurônio de saída que obtiver o maior valor final indicará a classe da amostra que está sendo analisada. Nesse trabalho é utilizada a rede MLP [Kruse *et al.* 2013], que é um modelo clássico de rede neural e bastante validado na literatura.

Figura 3 - Estrutura de uma rede neural.



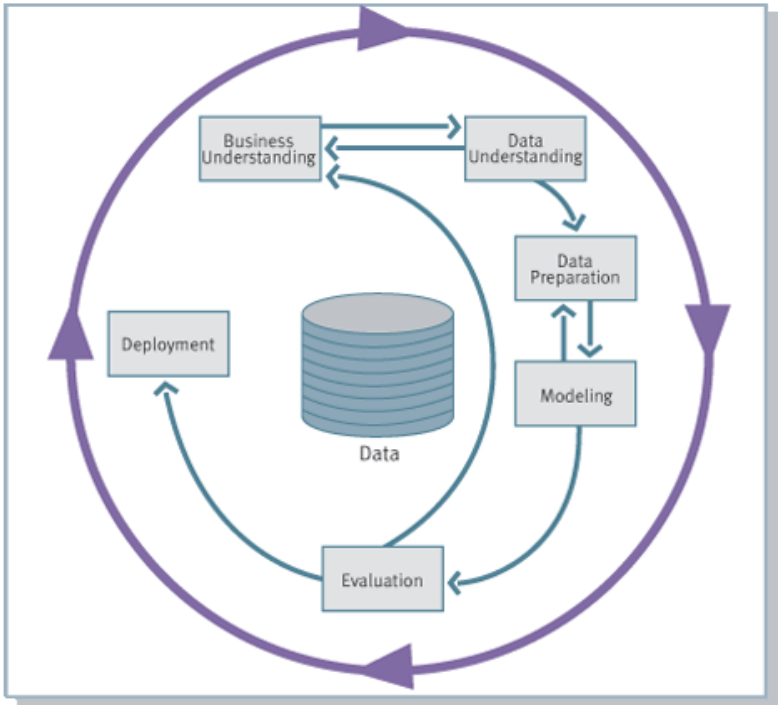
Fonte: Rolim *et al.* 2014.

2.2 PRÉ-PROCESSAMENTO

Podemos dizer que a mineração de texto é uma especialização da mineração de dados [Rezende *et al.* 2011]. A grande diferença entre a duas é o objeto de estudo, enquanto a mineração de dados trabalha com dados estruturados, a mineração de texto trabalha com dados não-estruturados, no caso texto. A mineração de texto tem por objetivo descobrir a maior quantidade de informação possível, diante de uma coleção de dados, por meio da identificação dos padrões e relações entre os dados [Capobianco 2015].

Para poder extrair informações relevantes dos dados, e para identificar os padrões neles contidos, é preciso executar algumas etapas, podemos observar um exemplo dessas etapas na figura 4, que vem mostrando o modelo da CRISP-DM (*Cross Industry Standard Process for Data Mining*).

Figura 4 - Fases do CRISP-DM.



Fonte: SPSS 2000.

A fase de preparação dos dados (*Data preparation*) ou pré-processamento dos dados, é altamente relevante para este trabalho, visto a grande quantidade de postagens contidas em nossa base de dados.

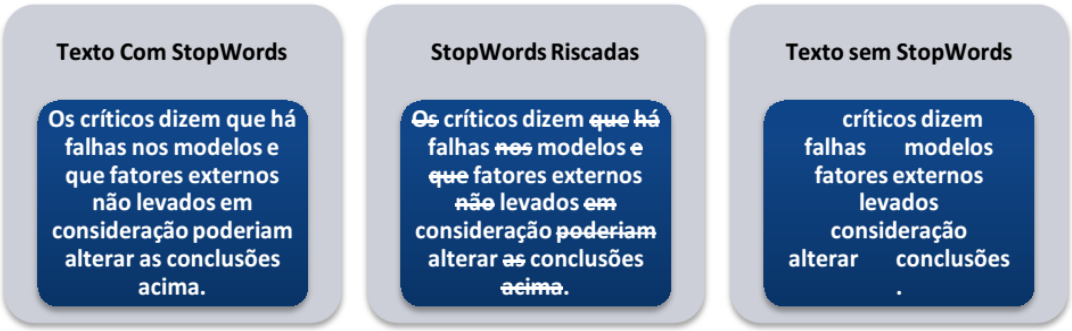
O pré-processamento é a etapa que é iniciada após a criação ou aquisição da base de dados, é responsável por tratar os dados, de forma que possam ser melhor trabalhados, e para que as técnicas computacionais possam ser aplicadas.

Para a etapa de pré-processamento deste trabalho, três técnicas foram empregadas: tokenização, remoção de *stopwords* e *stemming* [Hotho *et al.* 2005]. A tokenização é um processo necessário, ela “quebra” o texto de um documento, removendo as pontuações, e substituindo os elementos não textuais, como quebra de linha e tabulação, por espaço um espaço em branco simples.

As *stopwords* são as palavras que pouco acrescentam à representatividade da coleção de dados, ou que sozinhas nada significam. Exemplos de *stopwords* são palavras como artigos, pronomes e advérbios. É comum também remover pontuação e caracteres especiais. O conjunto de *stopwords* é chamado de *stoplist*. Essa eliminação de *stopwords* reduz significativamente a quantidade de termos,

diminuindo o custo computacional das próximas etapas [Rezende *et al.* 2011]. A figura 5 mostra um exemplo da aplicação desta técnica.

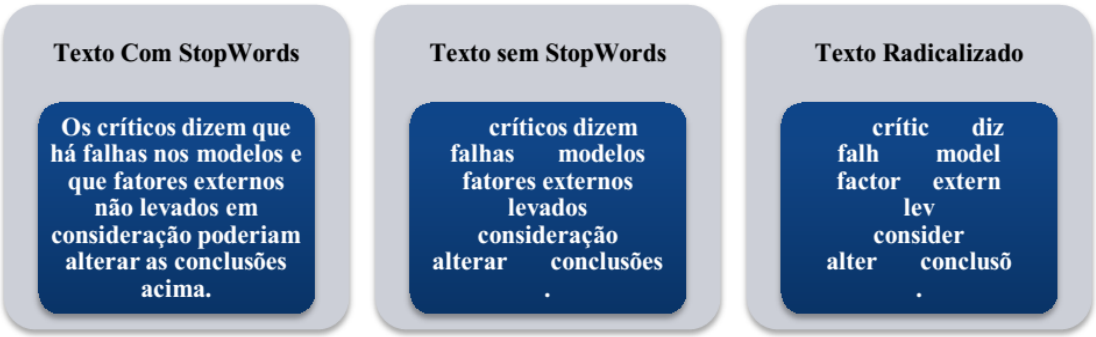
Figura 5 - Exemplo de remoção de *stopwords*.



Fonte: http://www.devmedia.com.br/mineracao-de-texto-analise-comparativa-de-algoritmos-revista-sql-magazine-138/34013

O *stemming* [Hotho *et al.* 2005] é uma técnica que reescreve as palavras do texto na sua forma básica. A palavra *stemming* deriva da palavra inglesa *stem*, que significa basicamente radical, ou seja, esta técnica transforma a palavra na sua forma radical, removendo sufixos. Podemos observar um exemplo da aplicação da técnica *stemming* na figura 6.

Figura 6 - Exemplo da técnica *stemming.*



Fonte: http://www.devmedia.com.br/mineracao-de-texto-analise-comparativa-de-algoritmos-revista-sql-magazine-138/34013

Além destas técnicas, também foi realizada a remoção de termos que tenham uma baixa frequência absoluta em relação à base de dados. Essa remoção é feita porque os termos com baixa frequência normalmente têm pouca influência na identificação das classes definidas.

2.3 RECUPERAÇÃO DE INFORMAÇÃO

Recuperação de informação (RI) é uma subárea da ciência da computação que tem por objetivo recuperar dados, informações ou até mesmo documentos de uma determinada coleção de dados. Baeza-Yates e Ribeiro-Neto [Baeza-Yates e Ribeiro-Neto 2013] definem o objetivo principal de um sistema de RI como sendo a recuperação de todos os documentos que são relevantes à necessidade de informação do usuário. Grande parte da dificuldade encontrada nos sistemas de RI está em saber como extrair a informação dos documentos e verificar a relevância daquela informação.

2.3.1 TF-IDF

É a combinação de duas técnicas estatísticas, o TF (*Term Frequency*) e IDF (*Inverse Document Frequency*). Essa medida é frequentemente utilizada na mineração de texto.

O TF mede o quão frequente um termo ocorre em um documento. Pode ser representado pela divisão entre a quantidade de vezes que um termo aparece no documento (frequenciaT) e o total de termos que o documento possui (N). Podemos ver a definição de TF na equação 6.

$$tf = \frac{frequenciaT}{N} \qquad (6)$$

O IDF mede quanto um termo é importante. Visto que o TF pode dar importância para termos como “de”, “para”, “como” por serem termos bastante comuns, aparecem muitas vezes, porém possuem pouca importância. O IDF procura dar um peso maior para termos que ocorrem mais raramente. Pode ser representado pela equação 7, onde ‘N’ é a quantidade total de documentos, e ‘n’ é a quantidade de documentos que possui um determinado termo.

$$idf = log(\frac{N}{n}) \qquad (7)$$

Logo uma medida razoável da importância de um termo pode ser obtida utilizando o produto da frequência do termo (TF) com a frequência inversa do documento (IDF) [Salton e Buckley 1988]. A fórmula obtida é representada por:

$$tfidf = tf \cdot idf \quad (8)$$

2.4 WEB 2.0

O criador do conceito da web 2.0 Tim O'Reilly o define como a mudança para a internet como plataforma, ressalta a importância da criação de aplicativos que se aproveitem do efeito de rede, para que se tornem melhores à medida que forem mais utilizados, de forma que aproveitem a inteligência coletiva [O'Reilly 2005].

Uma mudança importante que a web 2.0 trouxe foi a possibilidade de o usuário participar da criação de conteúdo. Diante disso, houve o surgimento de *softwares* que possibilitassem isso acontecer, exemplos destes softwares são as wikis, blogs, redes sociais e fóruns.

2.4.1 Fórum

O fórum de discussão é uma ferramenta de comunicação assíncrona, que tem como função principal promover discussões entre os participantes acerca de algum tema. Essa ferramenta é frequentemente utilizada nos ambientes virtuais de aprendizagem, por ser um recurso que proporciona a interação coletiva dos participantes.

Quando utilizado no meio educacional, fornece aos alunos um canal de comunicação onde podem ser expressas suas dúvidas, opiniões e respostas aos questionamentos existentes sobre algum assunto. Por isso comumente o fórum educacional possui professores ou tutores, como mediadores das interações, para que se possa auxiliar os alunos no processo de aprendizagem [Batista e Gobara 2007].

2.4.2 Wiki

A ferramenta wiki consiste em um gerenciador de conteúdo, os seus usuários podem criar e editar páginas dos mais variados temas, de forma colaborativa. Assim sendo, os usuários podem assumir diferentes papéis, como leitor, editor ou autor.

O wiki pode servir para diversos propósitos, tanto como repositório de informações, quanto ferramenta de apoio pedagógico [Ferreira *et al.* 2009]. Um exemplo bastante conhecido desse tipo de ferramenta é o *site* Wikipédia (https://www.wikipedia.org/), ferramenta na qual todo conteúdo existente é criado de forma colaborativa por usuários do mundo inteiro.

O Wikipédia possui diversas funcionalidades. Uma dessas funcionalidades é a possibilidade de colocar *links* no corpo do texto, que redirecionem para o assunto específico. Outra funcionalidade é a de poder destacar em negrito termos que possuem uma relevância maior no tema. Podemos observar na figura 7 alguns exemplos dessas funcionalidades.

Figura 7 - Página do Wikipédia, com *hiperlinks* e palavras em negrito em destaque.

Fonte: O autor.

3. REVISÃO DE LITERATURA

Esta seção contém os trabalhos relacionados a este trabalho. Eles mostram como a mineração de texto pode ser utilizada para auxiliar no meio educacional, especificamente fóruns educacionais. Também mostra a importância e a necessidade dessas ferramentas nesse meio.

3.1 TRABALHOS RELACIONADOS

Gerosa *et. al.* (2003) apresenta o AulaNet, um ambiente de interação entre alunos e professores, e demonstra o encadeamento e a organização das postagens através de hierarquia. Essa hierarquia das postagens é derivada da categorização de cada uma delas, as categorias podem ser: seminário (para uma mensagem raiz de uma discussão), questão (para propor um tópico para discussão), argumentação (para responder às questões, fornecendo um ponto de vista), contra argumentação (para oferecer uma posição contrária a uma argumentação), esclarecimento (para solicitar ou esclarecer dúvidas sobre uma mensagem). Esta categorização, implementada pelo próprio AulaNet, ajuda a observar os relacionamentos entre as postagens, dando assim subsídios aos professores para que possam coordenar eficazmente essas discussões, e avaliar o desenvolvimento da turma.

Podemos observar no trabalho de Ravi e Kim (2007) uma forma de identificar as interações dos alunos nos fóruns educacionais. Ele classifica individualmente as postagens dos alunos como “atos de discurso”, que podem ser: complemento (complemento de uma mensagem anterior), informação (informação, comando ou anúncio), correção (correção ou objeção a uma mensagem anterior), elaboração (elaboração de uma mensagem anterior ou descrição, incluindo elaboração de perguntas e respostas), pergunta (uma questão sobre um problema, incluindo questão sobre uma mensagem anterior), resposta (resposta a uma questão anterior, ou sugestão). Para a categorização são utilizadas técnicas de análise, *N-grams* e SVM (*Support Vector Machine*) para encontrar postagens que sejam categorizadas como pergunta ou resposta, e assim poder auxiliar o professor em encontrar

perguntas sem respostas, proporcionando uma gestão mais eficaz das postagens e dúvidas dos alunos.

Lin *et al.* (2009) propõe um sistema de classificação de gênero das postagens dos fóruns de discussão em ambientes educacionais. Utilizando a frequência das palavras como características para o sistema, e aplicando árvore de decisão para classificação. Os gêneros que este sistema tenta identificar são: anúncios (esclarece dúvidas dos outros), perguntas (dado uma informação desconhecida, propõe uma questão), interpretação (interpretação baseada nos fatos e ideias expostas, fazendo predições, analises e reflexões), conflito (opinião conflitante), afirmação (mantendo e defendendo a ideia discordada por outros), e outros (mensagens variadas que são difíceis de categorizar).

Shi *et al.* (2009) faz análise de sentimento baseado em tópicos de fóruns para agrupar postagens de temas relacionados. O principal objetivo deste trabalho é distinguir opiniões diferentes para um tema usando os conteúdos de fóruns. Para isso os autores transformam as postagens do fórum em uma lista de palavras e usa a frequência das mesmas para realizar o agrupamento.

Seguindo a mesma linha do trabalho citado acima, Li e Wu (2010) usam técnicas de agrupamento e classificação de texto para identificar fóruns que estão sendo muito acessados e para agrupar suas postagens por tópicos. Mais uma vez os atributos usados para aplicar as técnicas de mineração de texto é a frequência das palavras nos fóruns.

Oliveira Júnior *et. al.* (2011) apresenta uma ferramenta de classificação automática de postagens em fóruns educacionais. As postagens podem ser classificadas como positivas ou negativas, utilizando algoritmos de aprendizado de máquina, como *Naive Bayes* e SVM (*Support Vector Machine*). O algoritmo SVM apresenta resultados na avaliação, utilizando as métricas: *recall, precision, F-measure, p*orcentagem de mensagens classificadas corretamente e incorretamente, estatística Kappa, erro absoluto médio e raiz do erro quadrático médio. Essa classificação das postagens como positivas (expressam respostas, comentários pertinentes entre outros) e negativas (expressam dúvidas, conteúdo indevido entre outros) auxiliam o professor no momento de fornecer uma atenção maior a um determinado assunto.

Moghaddam e Ester (2011) revelam aspectos interessantes sobre análise de sentimento aplicada para sistemas de perguntas e respostas e analise de opinião. Eles mostram como fazer a mineração de opinião, e a partir dessa mineração tirar conclusões sobre as respostas dos tópicos criados. Para isso o sistema utiliza técnicas de mineração de texto para classificar se a postagem contém uma opinião positiva ou negativa.

O trabalho de Batista e Godara (2007) ressalta a importância da interação entre alunos e professores no fórum educacional. Revela a dificuldade de uma interação eficiente entre os atores envolvidos, destacando que grande parte dos professores utilizam predominantemente a ferramenta do fórum, para postagem de atividades. Essa afirmação pode ser comprovada por meio de uma entrevista realizada com professores e alunos do curso de pós-graduação *lato sensu* Orientação Pedagógica em Educação a Distância, de uma instituição pública de ensino superior. Essa entrevista mostrou que 80% dos professores usavam o fórum para postar atividades e dúvidas. Pôde-se constatar que 60% dos professores usavam o fórum para dar uma resposta a todos os participantes, enquanto apenas 50% dava uma resposta individualizada ao aluno.

Existem algumas diferenças entre os trabalhos citados e o nosso trabalho, podemos ver essas diferenças na tabela 1.

Todas essas pesquisas mostram a necessidade e a importância de ferramentas que auxiliem o professor na interação e acompanhamento dos alunos no fórum educacional. Porém identificamos que o nosso trabalho se difere dos demais nos seguintes pontos:

I. classificação das postagens em 3 categorias (dúvida, resposta e comentário neutro), identificando a dúvida do aluno;
II. extração do assunto da postagem logo após a classificação da mesma;
III. recomendação de um material de estudo auxiliar.

Tabela 1 - Tabela das diferenças entre os trabalhos.

Trabalhos	Identificação automática de dúvidas	Extração do assunto	Recomendação de material de estudo
Gerosa et al.	Não	Não	Não
Ravi e Kim	Sim	Não	Não
Lin et al.	Não	Não	Não
Shi et al.	Não	Não	Não
Li e Wu	Não	Não	Não
Oliveira Junior et al.	Sim	Não	Não
Moghaddam e Ester	Não	Não	Não
Batista e Godara	Não	Não	Não
Proposta	Sim	Sim	Sim

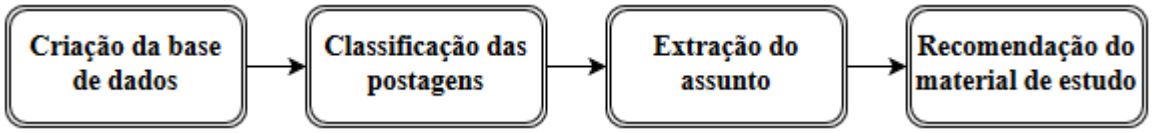
Fonte: O autor.

4. DESENVOLVIMENTO

Este capítulo aborda o processo de desenvolvimento deste trabalho, fornecendo as informações necessárias para compreensão de como ele foi planejado e do seu funcionamento. O Projeto divide-se em quatro etapas, e essa divisão pode ser observada na figura 8:

- Construção da base de dados utilizada no projeto.
- Classificador automático de postagens, utilizando aprendizagem de máquina.
- Extrator de assunto de postagens de fóruns educacionais, usando cálculo dos pesos de cada postagem.
- Sistema de recomendação que usa a resposta fornecida pelo extrator, recomenda vídeos do Youtube (https://www.youtube.com/).

Figura 8 - Fluxo completo do desenvolvimento.



Fonte: O autor.

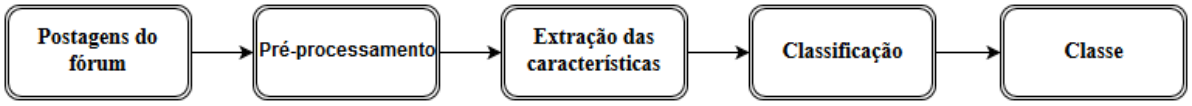
Mais detalhes sobre cada etapa serão apresentados nas seções subsequentes.

4.1 CLASSIFICAÇÃO DAS POSTAGENS

Esta primeira etapa do sistema é um classificador automático de postagens dos fóruns educacionais. As postagens podem ser classificadas como: Dúvida (postagem que contenha algum tipo de dúvida sobre o assunto do fórum); Neutra (postagem que não acrescenta a discussão, podendo ser apenas um comentário); Resposta (postagem que responde a algum questionamento, sendo ele proposto pelo professor, ou uma dúvida expressa por outro aluno).

Antes de que o classificador indique de qual classe uma determinada postagem pertence, existem alguns procedimentos que devem ser realizados. O fluxo do desenvolvimento desta fase é apresentado na figura 9.

Figura 9 - Fluxo da etapa de classificação das postagens.



Fonte: O autor.

4.1.1 Pré-processamento

Como técnica de pré-processamento aplicamos a tokenização, remoção de *stopwords* e *stemming* [Hotho *et al.* 2005]. A biblioteca utilizada para realizar a remoção dos *stopwords* e *stemming* foi a NLTK (http://www.nltk.org/). Essas técnicas foram aplicadas nas postagens, pois: a remoção de *stopwords* eliminará palavras com pouco valor no texto, que por ventura podem possuir um TF alto, porém pouca importância no contexto geral; o *stemming* colocará as palavras no seu radical, fazendo com que o TF de palavras realmente importantes aumente.

Deixando a parte a tokenização que foi utilizada em todos os casos, as demais técnicas de pré-processamento (remoção de *stopwords* e *stemming*) não foram definidas como regra, ou seja, em alguns testes utilizamos apenas uma técnica ou até mesmo nenhuma das duas. Essas variações foram necessárias para realizar comparações e analisar qual o melhor caso para este projeto, essas comparações são mostradas na seção 5.3.

4.1.2 Extração das características

Após o pré-processamento, é realizada a extração das características que serão os atributos para as técnicas de classificação. Foram definidos dois conjuntos de características. Entre as características utilizadas do primeiro conjunto estão:

- frequência das palavras da classe Dúvida;
- frequência das palavras da classe Neutra;
- frequência das palavras da classe Resposta;
- número de interrogações.

A extração dessas características é feita contando quantas palavras de cada classe aparecem no texto, para cada postagem.

Assim como no pré-processamento, fizemos testes usando conjuntos de características diferentes, sempre em busca do melhor resultado. As características utilizadas do segundo conjunto são:

- somatório do TF-IDF das palavras pertencentes a classe Dúvida;
- somatório do TF-IDF das palavras pertencentes a classe Neutra;
- somatório do TF-IDF das palavras pertencentes a classe Resposta;
- número de interrogações.

O TF-IDF de cada palavra é calculado e multiplicado pelo número de aparições da palavra, antes do somatório. Uma vez obtidos esses atributos, eles são disponibilizados como entrada para as técnicas de classificação.

A criação desse segundo conjunto é necessária, para que todos os cenários possíveis possam ser explorados. A escolha do TF-IDF nesse segundo conjunto, se deve a sua validação na literatura e por ser uma métrica de peso dos termos bastante utilizada na classificação de texto [Thorsten 1996, Zhang *et al.* 2011, Jing *et al.* 2002]

4.1.3 Classificação

Para essa pesquisa foi utilizada a ferramenta Weka (*Waikato Environment for Knowledge Analysis*) [Hall *et al.* 2009][Frank *et al.* 2010], versão 3.6.13. O Weka é uma ferramenta que disponibiliza uma coleção de algoritmos de aprendizado de máquina para utilização na mineração de dados. Nele estão implementadas as técnicas utilizadas no trabalho, assim como outras técnicas de aprendizagem de máquina.

Todas as técnicas implementadas utilizam uma etapa de treinamento e validação. No processo de avaliação aplicamos o método de *cross-validation* (Validação cruzada) [Arlot e Celisse 2010], com o número de *folds* igual a 10. Esse método divide os dados em vários grupos, dividindo-os entre treinamento e testes. Foram utilizados os seguintes algoritmos disponíveis no Weka: classificador

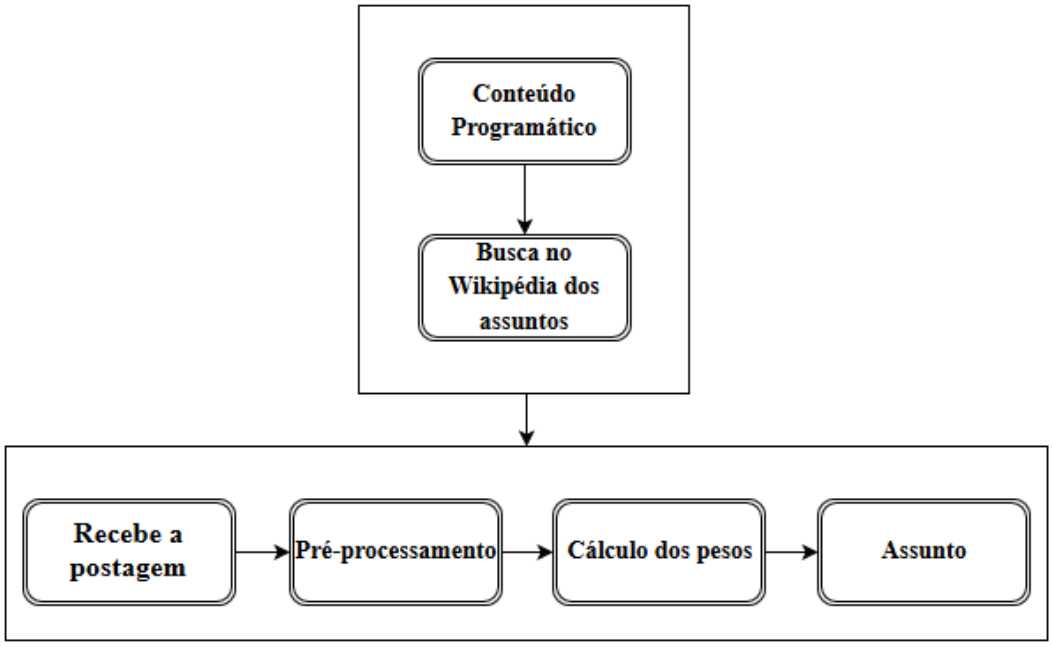
NaiveBayes, J48 e MultilayerPerceptron. Os parâmetros padrão da ferramenta não foram alterados.

Foram escolhidas essas técnicas de classificação, pois precisávamos de técnicas de naturezas diferentes, validadas pela literatura e fossem também comumente utilizadas para classificação textual.

4.2 EXTRAÇÃO DO ASSUNTO

Esta é a segunda etapa do sistema proposto, ela é responsável por identificar qual o assunto da dúvida contida na postagem do aluno. Logo após a classificação, sendo a postagem classificada como dúvida, ela entra na etapa de extração. O fluxo dessa etapa do projeto é melhor compreendido se observarmos a figura 10.

Figura 10 - Fluxo da etapa de extração de assunto.



Fonte: O autor.

Esta etapa é dividida em duas partes, a primeira parte recebe como entrada o conteúdo programático de uma disciplina provido pelo professor da mesma. Logo após receber essa entrada, todos os assuntos descritos no conteúdo programático são buscados no Wikipédia, e um banco de palavras é criado para cada um dos assuntos, cada palavra possui um peso associado.

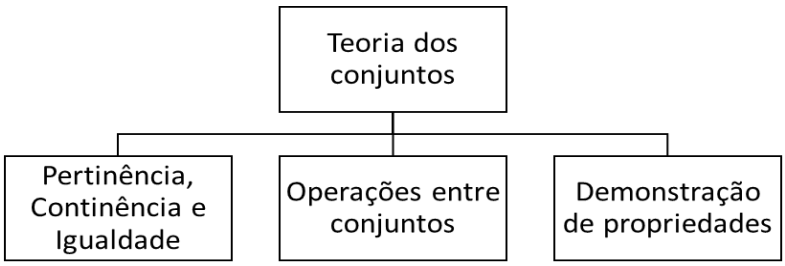
Em seguida, a postagem que anteriormente foi classificada, agora é novamente é submetida às técnicas de pré-processamento. A postagem então é comparada com os bancos de palavras, e verifica-se qual tem maior valor pelo somatório dos pesos das palavras. O assunto em que o somatório dos pesos for maior, é definido como assunto da postagem.

4.2.1 Busca dos assuntos no Wikipédia

Quando a disciplina é cadastrada no AVA, é necessário também que seja adicionado o plano de trabalho da disciplina, que aborda várias informações sobre ela. Dentre essas informações, existe o conteúdo programático, nele estão todos os assuntos que serão vistos no decorrer da disciplina.

Foi criado uma função que reconhece e extrai o conteúdo programático do plano de trabalho da disciplina. Em seguida outra função ainda organiza hierarquicamente os assuntos, com assuntos ‘pai’ e assuntos ‘filho’, como os próprios nomes sugerem, os assuntos indicados como ‘filho’ são derivações do assunto ‘pai’. A figura 11 mostra um exemplo de hierarquia dos assuntos.

Figura 11 - Exemplo da hierarquia dos assuntos.



Fonte: O autor.

Como podemos observar na figura 11, o assunto “Teoria dos conjuntos” possui um nível hierárquico maior (assunto “pai”), e os outros no nível abaixo, são os assuntos “filho”.

Para cada assunto, seja ele ‘pai’ ou ‘filho’, é feita uma consulta no Wikipédia de cada um deles. Todos os *hiperlinks* e as palavras em negrito são adicionados no banco de termos de seu respectivo assunto, juntamente com seu peso. O peso é calculado pela equação 6 (TF), que é razão entre a quantidade de aparições do termo e a quantidade total de termos no texto do Wikipédia.

4.2.2 Pré-processamento

Como na etapa de classificação, foram utilizadas as técnicas de remoção de *stopwords* e o *stemming*. Cada postagem é submetida à ambas as técnicas, assim como os termos adicionados no banco de termos de cada assunto.

4.2.3 Cálculo dos pesos

Assim como explicado anteriormente, os pesos dos termos correspondentes a cada assunto, é a razão entre a quantidade de aparições do termo e a quantidade total de termos no texto do Wikipédia.

Com todos os bancos de termos preenchidos, e cada termo com seu respectivo peso, é possível então definir o assunto de uma postagem. Para isso cada palavra contida na postagem é comparada com os bancos. Para cada termo da postagem que esteja contido no banco de termos, o peso do termo do banco é somado. Ao termino de cada banco o somatório é armazenado e zerado, para que o próximo somatório possa iniciar.

O somatório do banco de termos que possuir maior valor, é o assunto da postagem.

4.3 RECOMENDAÇÃO DE MATERIAL DE ESTUDO

Esta é a última etapa do sistema, nesta etapa, é dado como entrada o assunto extraído da postagem assim como visto na etapa anterior, e é buscado um vídeo do Youtube que melhor atenda o assunto. O vídeo encontrado é sugerido ao aluno.

Após a etapa de extração do assunto da postagem, o assunto é passado como argumento para a função responsável por buscar vídeos no Youtube. Essa função retorna o link do vídeo melhor ranqueado, o algoritmo do Youtube por padrão ordena a lista de vídeos por relevância, logo, é suposto que seja o melhor para a pesquisa realizada. Ao obter este link, ele é recomendado ao aluno, servindo assim como auxílio para que o aluno possa sanar sua dúvida.

5. RESULTADOS

Esta seção apresenta as bases de dados utilizadas no experimento, as métricas de avaliação e os resultados obtidos nos testes realizados nas três etapas: classificação da postagem, extração do assunto da postagem, e recomendação de material de estudo.

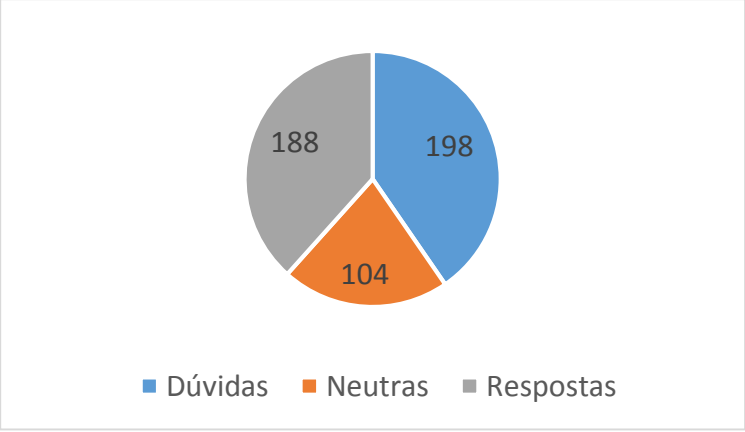
5.1 BASES DE DADOS

Esta etapa do trabalho diz respeito a criação da base de dados utilizadas neste projeto. Durante o processo de desenvolvimento desta fase, de classificação, do projeto foram utilizadas 3 bases de dados diferentes.

A base de dados 1 foi retirada do AVA da UFAL (Universidade Federal de Alagoas), e as postagens foram retiradas de várias disciplinas diferentes do curso à distância de bacharelado em sistemas de informação. Já a base de dados 2, foi retirada também do AVA da UFAL, porém, das disciplinas de algoritmos e estruturas de dados I e II, do mesmo curso.

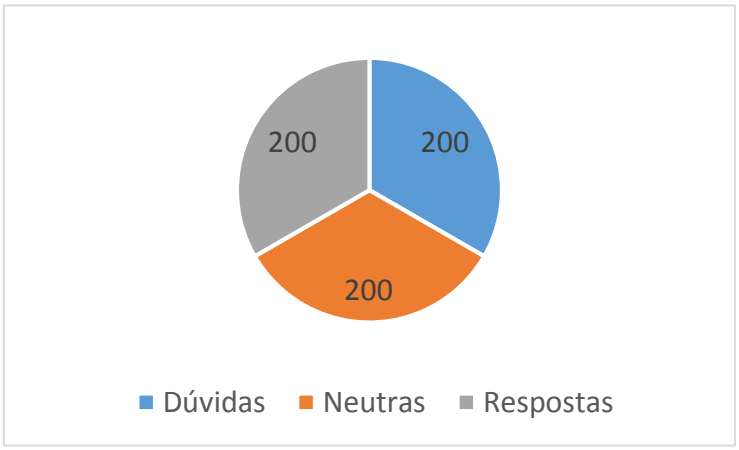
As figuras 12 e 13 mostram a distribuição das classes nas bases de dados utilizadas em toda fase de classificação do projeto, onde os valores expostos representam as respectivas quantidades de postagens pertencentes às classes indicadas.

Figura 12 - Distribuição da base de dados 1.



Fonte: O autor.

Figura 13 - Distribuição da base de dados 2.



Fonte: O autor.

A terceira base de dados (base de dados 3) possui 387 postagens, e diferentemente das outras duas, não está distribuída conforme as classes Dúvida, Neutra ou Resposta. A distribuição desta base é feita conforme o assunto ao qual pertencem. Podemos ver a distribuição na figura 14.

Figura 14 - Distribuição da base de dados 3.

Fonte: O autor.

Esta base de dados foi desenvolvida na disciplina de tópicos avançados em inteligência artificial, do curso de bacharelado em ciência da computação da UFRPE. Os alunos dessa disciplina tiveram que escrever postagens que apresentassem

dúvidas ou respostas com relação aos assuntos da disciplina de matemática discreta. Ao invés de separar as postagens por classe (Dúvida, Neutra e Resposta), a base foi distribuída segundo o assunto que representavam na disciplina proposta. Os valores apresentados na figura 14 representam a quantidade de postagens que os assuntos contem.

5.2 MÉTRICAS DE AVALIAÇÃO

Neste trabalho foram utilizadas as seguintes métricas de análise: Precisão, Cobertura e Medida-F. Segundo Friedman *et al.* [Friedman *et al.* 1997], a precisão avalia o quanto o sistema acerta. A métrica de precisão aplicada no trabalho é baseada na quantidade de postagens classificadas corretamente. A precisão é calculada pela Equação 9.

$$Precisao = \frac{tp}{tp + fp}, \qquad (9)$$

Onde *tp* (*true positive*) representa o número de verdadeiros positivos, enquanto *fp* (*false positive*) é o número de falsos positivos. *True positive* são instâncias que pertencem a uma determinada classe, e que foi classificada corretamente. *False positive* são instância que foram preditas como pertencentes a uma classe, porém incorretamente.

A métrica de Cobertura [Roncero 2010] avalia a porcentagem de quantos itens relevantes foram de fato classificados como relevantes. Podemos observar sua equação 10 logo abaixo:

$$cobertura = \frac{tp}{tp + fn}, \qquad (10)$$

onde *tp* representa o número de verdadeiros positivos e *fn* (*false negative*) é o número de falsos negativo. *False negative* são instâncias que deveriam pertencer a uma classe, porém foram classificadas como pertencente a outra.

A métrica Medida-F pode ser interpretada como uma média ponderada da precisão e da cobertura, onde o F tem o seu melhor valor em 1 e o pior em 0, conforme pode ser notado na equação 11.

$$F = 2 \cdot \frac{precisao \cdot cobertura}{precisao + cobertura} \quad (11)$$

Outra métrica que foi utilizada foi a acurácia, que é definida como sendo a razão da soma de todos os verdadeiros positivos pela soma de todos os verdadeiros positivos e falsos positivos para todas as classes [Barros *et al.* 2012].

5.3 CLASSIFICAÇÃO DAS POSTAGENS

A avaliação da etapa de classificação será feita usando a métrica medida-F. Quanto mais próximo a medida-F for de 1, melhor o resultado. As bases de dados utilizadas foram a base de dados 1 (BD1) e a base de dados 2 (BD2), apresentadas na seção 5.1. Como mencionado anteriormente, o ambiente experimental utilizado foi o programa Weka.

Primeiramente avaliamos qual seria a melhor técnica de classificação para a nossa abordagem. As técnicas utilizadas, com seus respectivos algoritmos contidos no Weka, foram: NaiveBayes, J48, MultilayerPerceptron. Como citado anteriormente, os parâmetros padrão do weka não foram alterados para o experimento.

Tabela 2 - Média das medidas-F das técnicas de classificação

Técnicas de classificação	*Média da Medida-F*	
	BD1	*BD2*
NaiveBayes	0.526	0.701
J48	0.885	0.882
MultilayerPerceptron	0.962	0.972

Fonte: O autor.

Para a realização deste primeiro teste, não foi aplicado nenhum tipo de pré-processamento nas postagens, as características selecionadas foram as frequências de palavras de cada classe (Dúvida, Neutra, Resposta), e o limite de corte foi deixado em 0. Podemos observar o resultado na tabela 2.

O resultado expresso na tabela 2, mostra que o algoritmo MultilayerPerceptron (MLP) é o mais adequado para resolução do problema abordado por este trabalho, apresentando um valor de 0,962 na média obtida da medida-F das três classes na BD1 e 0,972 na BD2. Este resultado ratifica a conclusão de Rolim *et al.* (2014), que aponta o MLP como melhor algoritmo de classificador para este o problema de classificação de postagens.

Com o melhor algoritmo definido, resta definir quais serão as características que devem ser selecionadas, quais técnicas de pré-processamento devem ser empregadas, e qual limite de corte deve ser estabelecido, a fim de que possamos encontrar o melhor cenário possível e aumentar ainda mais o valor da medida-F.

O limite de corte foi implantado para que palavras de baixa relevância fossem eliminadas da coleção de palavras. Por isso palavras com frequência ou TF-IDF menor que os valores estabelecidos são descartadas.

Dividiremos os testes em dois tipos, baseado nas características utilizadas. O primeiro tipo usa a frequência de palavras como característica, e o segundo usa o TF-IDF. Os dois tipos sofrerão variações de técnicas de pré-processamento e limites de corte.

Como o principal objetivo do sistema é encontrar dúvidas, o melhor resultado, será o que possuir o maior valor da medida-F da classe Dúvida.

5.3.1 Usando frequência das palavras

Para o primeiro cenário, as características selecionadas foram as frequências das palavras de cada classe. Não foi aplicado nenhum tipo de pré-processamento, e a única variação que ocorreu, foi do limite de corte, que variou entre 0, 5 e 10. Os limites de cortes utilizados foram definidos com base nos limites utilizados em Rolim *et al.* (2014). O resultado deste primeiro cenário se encontra na tabela 3.

Tabela 3 - Classificação sem usar tecnicas de pré-processamento (frequência).

Base de dados	Limite de corte em 0			Limite de corte em 5			Limite de corte em 10		
	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta
BD1	**0.96**	0.935	0.979	0.738	0.708	0.774	0.728	0.623	0.745

Base de dados	Limite de corte em 0			Limite de corte em 5			Limite de corte em 10		
	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta
BD2	**0.977**	0.961	0.977	0.915	0.844	0.835	0.821	0.797	0.827

Fonte: O autor.

No segundo cenário, a única diferença em comparação com o primeiro cenário, é a adição da técnica de pré-processamento, no caso, a remoção de *stopwords.* Vemos o resultado na tabela 4.

Tabela 4 - Classificação usando remoção de *stopwords* (frequência).

Base de dados	Limite de corte em 0			Limite de corte em 5			Limite de corte em 10		
	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta
BD1	**0.963**	0.953	0.966	0.793	0.768	0.801	0.736	0.634	0.71
BD2	**0.962**	0.933	0.955	0.906	0.859	0.872	0.891	0.823	0.834

Fonte: O autor.

Assim como o primeiro cenário, o melhor resultado se apresentou quando o limite de corte estava em 0, aumentando o valor no BD1 para 0.963 e reduzindo o valor na BD2 para 0.962.

Para o terceiro cenário, adicionalmente ao segundo cenário, aplicou-se mais uma técnica de pré-processamento, o *stemming.* O resultado está expresso na tabela 5.

Tabela 5 - Classificação usando remoção de *stopwords* e *stemming* (frequência).

Base de dados	Limite de corte em 0			Limite de corte em 5			Limite de corte em 10		
	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta
BD1	**0.892**	0.825	0.922	0.779	0.777	0.806	0.77	0.713	0.782
BD2	0.872	0.874	0.895	**0.879**	0.809	0.854	0.842	0.785	0.796

Fonte: O autor.

Podemos notar que o melhor resultado permaneceu no caso em que o limite de corte estava em 0 na BD1, com o valor de 0.892. Diferentemente do resultado da

BD1, e dos resultados anteriores, o melhor resultado para a BD2 foi quando o limite de corte estava definido em 5, no valor de 0.879.

Analisaremos mais a frente os resultados desta fase (usando frequência das palavras), juntamente com o resultado da fase usando TF-IDF, para enfim definir qual cenário mais adequado para resolução do problema abordado neste trabalho.

5.3.2 Usando TF-IDF

No primeiro cenário, as características selecionadas foram os somatórios dos TF-IDF das palavras de cada classe, assim como todos os demais cenários deste tópico serão. Não foi aplicado nenhum tipo de pré-processamento, e a única variação que ocorreu, foi do limite de corte, que variou entre 0, 0.00109 e 0.00247. Os limites de cortes utilizados foram definidos após o teste de vários outros valores, os valores adotados foram o que apresentaram os melhores resultados. O resultado deste primeiro cenário se encontra na tabela 6.

Tabela 6 - Classificação sem técnicas de pré-processamento (tf-idf).

Base de dados	Limite de corte em 0			Limite de corte em 0.00109			Limite de corte em 0.00247		
	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta
BD1	0.832	0.48	0.636	**0.854**	0.819	0.795	0.834	0.705	0.748
BD2	**0.938**	0.908	0.895	0.885	0.817	0.777	0.846	0.655	0.585

Fonte: O autor.

Observamos que o melhor resultado (valor da medida-F da classe Dúvida) se apresentou quando o limite de corte foi definido em 0.00109, no valor de 0.854 na BD1. E para o BD2 o melhor resultado foi quando o limite de corte estava em 0, apresentando o valor de 0.938.

No segundo cenário, diferente do primeiro, foi aplicada a técnica de remoção de *stopwords* como pré-processamento. Os valores de corte foram mantidos.

O melhor resultado tanto na BD1 quanto na BD2, se apresentou quando o limite de corte estava definido em 0.00109, nos valores de 0.911 e 0.941 respectivamente. Encontramos os resultados na tabela 7.

Tabela 7 - Classificação usando remoção de stopwords (tf-idf).

Base de dados	Limite de corte em 0			Limite de corte em 0.00109			Limite de corte em 0.00247		
	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta
BD1	0.832	0.879	0.78	**0.911**	0.898	0.87	0.855	0.763	0.788
BD2	0.85	0.549	0.536	**0.941**	0.899	0.895	0.893	0.822	0.8

Fonte: O autor.

Neste terceiro e último cenário, além de aplicar a remoção de *stopwords*, também será aplicado o *stemming*. Os limites de corte serão mantidos novamente. Os resultados deste terceiro cenário estão contidos na tabela 8.

Tabela 8 - Classificação usando remoção de *stopwords* e *stemming* (tf-idf).

Base de dados	Limite de corte em 0			Limite de corte em 0.00109			Limite de corte em 0.00247		
	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta
BD1	0.823	0.667	0.711	**0.908**	0.916	0.87	0.868	0.849	0.823
BD2	0.843	0.639	0.611	**0.926**	0.885	0.88	0.882	0.793	0.801

Fonte: O autor.

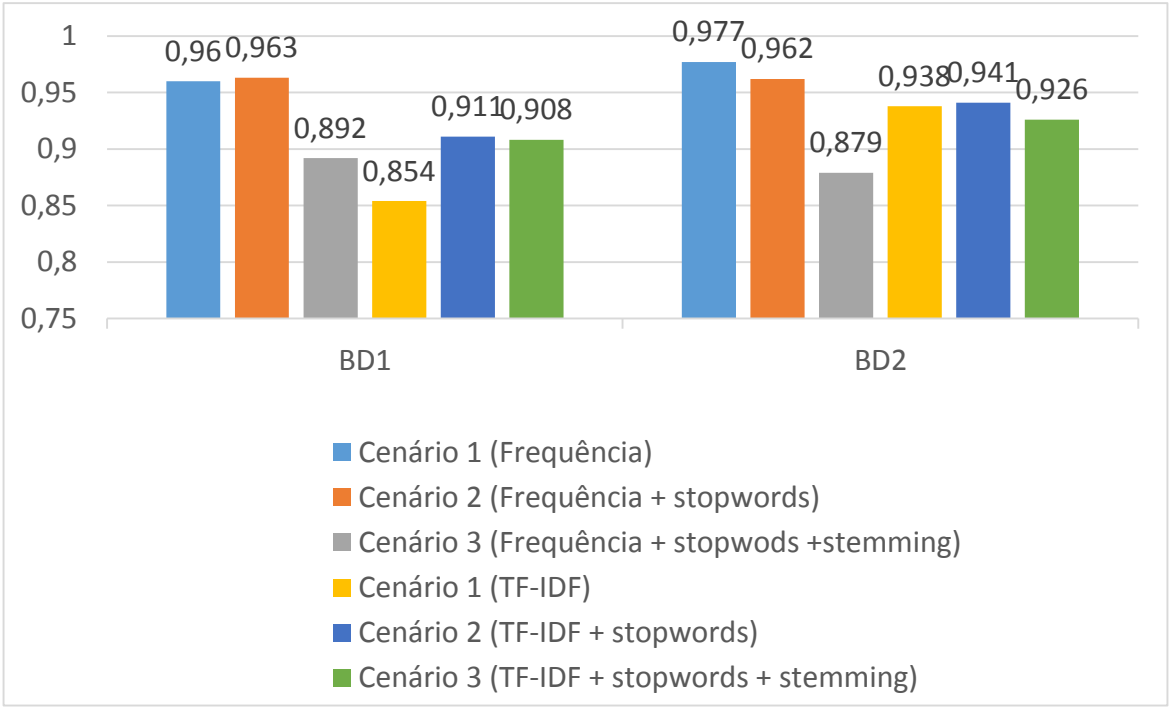
Assim como no cenário anterior, o melhor valor da medida-F se apresentou quando o limite de corte foi definido em 0.00109, no valor de 0.908 na BD1 e de 0.926 na BD2.

5.3.3 Analise dos resultados

Para definir qual é o melhor cenário para a classificação, foi escolhido como critério o valor da medida-F da classe Dúvida, visto que o principal objetivo deste trabalho é identificar a dúvida do aluno, para posteriormente recomendar um material de estudo auxiliar, a fim de sanar tal dúvida.

Ao observarmos os resultados apresentados, notamos claramente que os dois primeiros cenários (usando frequência de palavras) obtiveram os melhores resultados, tanto na BD1 quanto na BD2, com os valores de 0.96 e 0.963 para a BD1, e 0.977 e 0.962 para a BD2. A figura 15 nos mostra os maiores valores atingidos da medida-F da classe Dúvida para as duas bases de dados utilizadas.

Figura 15 – Maiores valores atingidos da medida-F da classe Dúvida.



Fonte: O autor.

Podemos observar que o cenário 1 e 2 são os que apresentam os maiores valores da medida-F.

O cenário 1 faz uso da frequência das palavras das classes como característica, e não faz uso de técnicas de pré-processamento. Os valores apresentados na figura 15 para o cenário 1, foram alcançados quando o limite de corte estava em 0.

O cenário 2, faz uso da frequência das palavras das classes, e faz a remoção de *stopwords* como técnica de pré-processamento. Assim como o cenário 1, os melhores valores foram atingidos quando o limite de corte estava em 0.

Se calcularmos a média dos valores atingidos na BD1 e BD2 do cenário 1 e 2, chegamos aos valores de 0.968 e 0.962 respectivamente. O valor da média do cenário 1 é levemente superior à média do cenário 2. Portanto, o cenário 1 é o mais adequado para ser utilizado na etapa de classificação das postagens, pois possui o maior valor da medida-F, o que representa uma maior acurácia.

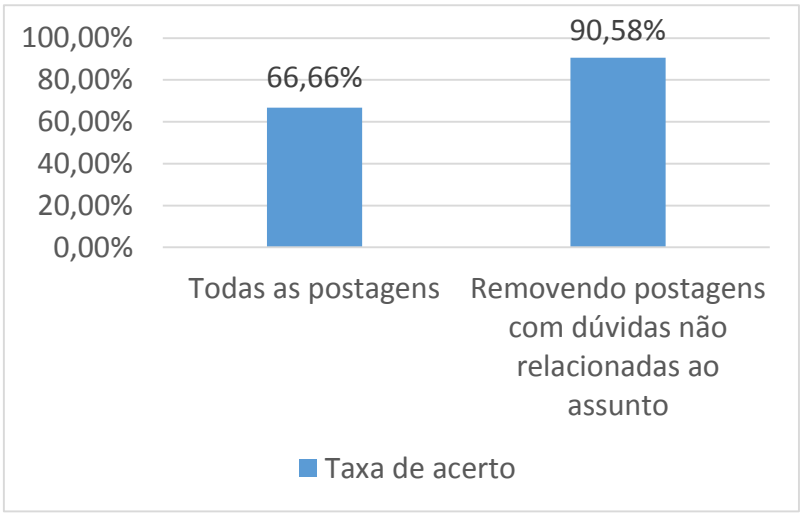
5.4 EXTRAÇÃO DO ASSUNTO

Como explicado na seção 4.2, foram utilizadas duas bases de dados (base de dados 2 e base de dados 3) com o intuito de avaliar a taxa de acerto do algoritmo de extração de assunto. Para os testes realizados, foi necessário também o conteúdo programático dos cursos das quais as bases estão relacionadas, no caso, Algoritmos e estrutura de dados para a base de dados 2, e matemática discreta para a base de dados 3. Os dois conteúdos programáticos utilizados podem ser encontrados nos anexos I e II.

A base de dados 2, é composta por 600 postagens, divididas em: 200 postagens classificadas como dúvida; 200 classificadas como neutra; 200 classificadas como resposta. No caso da extração foram utilizadas para teste, apenas as postagens classificadas como dúvida. Como técnicas de pré-processamento, foram aplicadas a remoção de *stopwords* e *stemming*, tanto nas postagens quanto nos termos relacionados a cada assunto.

Foi possível extrair corretamente o assunto de 66,66% das postagens. Contudo, identificamos que algumas postagens dificultam a extração do seu assunto corretamente, tais como: “Professor, qual seria a melhor forma para resolver a primeira questão da lista?”; “Que horas abre o laboratório?”; “Qual será a data da prova?”.

Figura 16 - Gráfico da diferença dos resultados da extração.



Fonte: O autor.

Esses tipos de postagens dificultam a extração do seu assunto, tendo em vista que para uma extração correta, precisamos que o assunto esteja relacionado com algum assunto no conteúdo programático da disciplina.

Realizamos outro teste removendo esses tipos de postagens, a taxa de acerto subiu para 90,58%. Este resultado mostra que existe a necessidade da otimização do algoritmo de extração, de forma que de alguma forma possa eliminar as postagens que não são relacionadas a algum assunto da ementa do curso. A diferença entre os dois resultados pode ser observada na figura 16.

Já a base de dados 3, é distribuída conforme o assunto à qual pertence: 119 postagens anotadas como “teoria dos conjuntos”; 38 anotadas como “teoria dos números”; 122 anotadas como “relações, funções e sequências”; 118 anotadas como “lógica”. Assim como na base de dados 2, foi aplicado a remoção de *stopwords* e *stemming* como técnicas de pré-processamento. O resultado deste teste pode ser observado na tabela 9.

Tabela 9 - Taxas de acerto dos assuntos da base de dados 3.

Assunto	*Taxa de acerto*
Teoria dos conjuntos	71,43%
Teoria dos números	84,21%
Relações, funções e sequências	62,3%
Lógica	86,44%
Média	76,1%

Fonte: O autor.

A média da taxa de acerto obtida foi de 76,1%. Analisando as partes que compõem este resultado, observamos que os resultados individuais não foram uniformes, um exemplo dessa diferença pode ser visto pela taxa de acerto de “Relações, funções e sequências”, 62,3%, e pela taxa de “Lógica” de 86,44%.

Isso nos mostra que mesmo com assuntos do mesmo universo, no caso, matemática discreta, houve uma diferença considerável entre os resultados. Esta diferença pode ser creditada a como o conteúdo programático da disciplina é descrito. Além das postagens já mencionadas anteriormente, que não possuem assuntos relacionados a algum assunto da ementa da disciplina.

Vale salientar que a base de dados 3 não possui postagens com comentários neutros, diferentemente da base de dados 2 que possui. Logo, esse fato pode ter tido alguma influência no valor de 76,1% apresentado na extração da base de dados 3.

Ao finalizarmos os testes, identificamos o desafio de melhorar a taxa de acerto da extração de assunto, mesmo obtendo taxa boas, de 66,66% para base de dados 2 e 76,1% para a base de dados 3. Logo, consideramos a taxa de acerto dependendo das variáveis de entrada, um percentual entre 66,66% e 76,1, observada a influência dessas variáveis no resultado final.

Para a etapa de recomendação não serão realizados os testes com o objetivo de obter uma taxa de acerto, visto que a recomendação do vídeo do Youtube está estreitamente relacionada ao assunto da postagem. Portanto, avaliar o desempenho do Youtube com relação à busca dos vídeos foge do escopo deste trabalho.

6. CONCLUSÃO E TRABALHOS FUTUROS

Com a grande disseminação da educação a distância os ambientes virtuais de aprendizagem estão sendo largamente usados. Nestes ambientes as ferramentas da Web 2.0, como fóruns, blogs, wiki, tem ganhado um papel no processo de ensino-aprendizagem. Os fóruns têm uma característica importante, nele os alunos postam dúvidas e possíveis respostas para questões levantas pelo professor.

Com isso, esta ferramenta produz um conteúdo bastante valioso para o curso que a utiliza. Contudo, devido à grande quantidade de alunos normalmente inscritos em cursos com plataformas online, torna-se difícil realizar o acompanhamento direcionado para cada aluno de forma. Além disto, o acesso ao conteúdo do fórum também não é utilizado de forma eficiente.

Diante dessas considerações podemos afirmar a importância deste trabalho, para que auxilie os professores no acompanhamento dos alunos, e que auxilie os alunos em busca das respostas aos seus questionamentos concernentes a um assunto específico.

A principal proposta deste trabalho, foi a criação de uma nova abordagem, para ajudar o professor a solucionar dúvidas dos alunos de forma automatizada, de forma que pudesse otimizar o processo de ensino-aprendizagem. Essa abordagem auxilia tanto o professor, reduzindo o tempo empregado para responder todos os questionamentos dos alunos, quanto o aluno, indicando materiais de estudo que possam auxiliá-lo na resolução da sua dúvida.

Entre as dificuldades encontradas no processo de desenvolvimento deste trabalho, podemos citar a criação das bases de dados utilizadas, visto que foram retiradas do *moodle* da UFAL de forma manual, tendo ainda que rotular cada postagem de acordo com seu conteúdo também manualmente. Outra dificuldade que pode ser mencionada, diz respeito a utilização de bibliotecas no desenvolvimento do classificador automático, ao tentar integrar essas bibliotecas (Scikit-learn, Orange, Python-weka) que reduziriam consideravelmente o tempo gasto na codificação dessa etapa, foram identificados problemas de compatibilidade com o ambiente de desenvolvimento.

6.1 TRABALHOS PUBLICADOS

Durante o processo de desenvolvimento deste trabalho, publicamos um artigo no Encontro nacional de inteligência artificial e computacional (ENIAC), no ano de 2014, juntamente com os professores Filipe Cordeiro e Rafael Ferreira, intitulado “Reconhecimento de padrões aplicados a comentários de fóruns educacionais”. Este artigo pode ser encontrado no anexo III.

Conseguimos a 2ª colocação no concurso de *posters* da semana da computação (SECOMP), realizada na UFRPE no ano de 2015, ao apresentarmos o *poster* intitulado “Análise comparativa entre técnicas de classificação de texto para identificação de dúvidas em fóruns educacionais”. O *poster* foi produzido juntamente com os professores Filipe Cordeiro e Rafael Ferreira. Este *poster* pode ser encontrado no anexo IV.

6.2 TRABALHOS FUTUROS

Como trabalhos futuros, iremos desenvolver uma forma de eliminar as postagens de dúvidas que não são relacionadas com algum assunto do conteúdo programático da disciplina, visto que como mencionado anteriormente, reduziram drasticamente o percentual de acerto do extrator de assunto. Também desenvolveremos um padrão para a descrição do conteúdo programático da disciplina, já que assim como o problema da postagem, contribuiu para a redução da taxa de acerto do extrator. Iremos realizar experimentos de extração de assuntos, utilizando outras fontes além do Wikipédia para coletar os termos relacionados.

Além disso, adicionaremos outras fontes para recomendação do material de estudo auxiliar, para somar aos vídeos do Youtube que já são recomendados.

REFERÊNCIAS

Akyuz, H. I., Kurt, M. *Effect of teacher's coaching in online discussion forums on students' perceived self-efficacy for the educational software development. Procedia - Social and Behavioral Sciences, World Conference on Learning, Teaching and Administration Papers, vol. 9, no. 0, pp. 633 – 637, 2010.* Disponível em*:* <http://www.sciencedirect.com/science/article/pii/S1877042810023141>

Arlot, S., Celisse, A., *et al. A survey of cross-validation procedures for model selection. Statistics surveys*, 4:40–79, 2010.

Azevedo, B. F. T., Behar, P. A., Reategui, E. B. Análise temática das mensagens de discussões online, Cadernos de Informática – Volume 6 – Número 1, 2011.

Baeza-Yates, R. e Ribeiro-Neto, B. Recuperação de informação: conceitos e tecnologia das máquinas de busca. Bookman editora, 2ª edição, 2013.

Baker, R. e Yacef, K. *The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

Barros, M. G. e Carvalho, A. B. G. As concepções de interatividade nos ambientes virtuais de aprendizagem. Tecnologias digitais na educação, 1(1):209–232, 2011.

Barros, M., de Morais Gomes, R., Alencar, M., Júnior, P., & Costa, A. Avaliação de classificação de tráfego ip baseado em aprendizagem de máquina restrita à arquitetura de serviços diferenciados. *Vol 1*, p. 10-20, 2012.

Bhargava, N., Sharma, G., Bhargava R., e Mathuria M., *Decision tree analysis on j48 algorithm for data mining. International Journal*, vol. 3, no. 6, 2013.

Batista, E. M. e Gobara, S. T. O fórum on-line e a interação em um curso a distância. IX Ciclo de Palestras sobre Novas Tecnologias na Educação, 2007. Disponível em: < http://www.cinted.ufrgs.br/ciclo9/artigos/8cErlinda.pdf>

Capobianco, K.R. Avaliação da etapa de pré-processamento na mineração de texto em redes sociais digitais. 40 p. Trabalho de Conclusão de Curso – Versão Pre-liminar (Bacharelado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina–PR, 2015.

Cheng, C. K., Par, D. E., Collimore, L.-M., and Joordens, S. *Assessing the effectiveness of a voluntary online discussion forum on improving students course performance. Computers & Education, 56(1):253 – 261, 2011. Serious Games.*

Dillenbourg, P. S. e Schneider D. *Virtual learning environments. Proceedings of the 3rd Hellenic Conference on Information & Communication Technologies in Education*, pp. 3–18, 2002.

Ferreira, A. A., Silva, B. D., e Siman, L. M. C. Web 2.0 e o ensino de história: trabalhando com wiki. ENCONTRO NACIONAL PERSPECTIVAS DO ENSINO DE HISTÓRIA, 7, Uberlândia, Minas Gerais, Brasil, 2009 – “Anais do VII Encontro Nacional “Perspectivas do Ensino de História.” Uberlândia : Universidade Federal de Uberlândia, 2009. ISBN 978-85-7078-218-2.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., e Trigg, L. *Weka-a machine learning workbench for data mining. Data Mining and Knowledge Discovery Handbook*, pp. 1269–1277, 2010.

Freitas, M. A. S. Avaliação da Aprendizagem em ambientes de formação online: aportes para uma abordagem hermenêutica. PhD thesis, UFBA: Faculdade de Educação, 2009.

Friedman, N., Geiger, D., e Goldszmidt, M. *Bayesian network classifiers. Machine learning*, 29(2-3):131–163, 1997.

Gerosa, M. A., Pimentel, M. G., Fucks, H. e Lucena, C. J. P. Coordenação de Fóruns Educacionais: Encadeamento e Categorização de Mensagens. XIV Simpósio Brasileiro de Informática na Educação, 2003.

Hall, M., Frank, E., Holmes, G., Pfahringer B., Reutemann P., e Witten I. H. *The weka data mining software: an update, ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

Haykin S. *Neural Networks: A Comprehensive Foundation*. 2ª Edição. Macmillan, Nova Iorque, 1999.

Hotho, A., Nurnberger, A., e Paas, G. *A brief survey of text mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 20(1):19–62, 2005.

Jing, L-P., Huang, H-K., Shi, H-B. *Improved feature selection approach TFIDF in text mining. In: Machine Learning and Cybernetics, 2002. Proceedings*. 2002 *International Conference on*. IEEE, 2002. p. 944-946.

Kim, J. *Influence of group size on students’ participation in online discussion forums. Computers & Education,* 62(0):123 – 129, 2013.

Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M., e Held, P. *Multilayer perceptrons. Computational Intelligence*, pp. 47–81, 2013.

Lin, F.-R., Hsieh, L.-S., Chuang, F.-T. *Discovering genres of online discussion threads via text mining. Computers & Education*, vol. 52, no. 2, pp. 481–495, 2009.

McCallum, A., e Nigam, K. "*A comparison of event models for naive bayes text classification*." *AAAI-98 workshop on learning for text categorization*. Vol. 752. 1998.

Mohamad, S. K. e Tasir, Z. *Educational data mining: A review. Procedia - Social and Behavioral Sciences, The 9th International Conference on Cognitive Science, vol. 97,*

no. 0, pp. 320 – 324, 2013.Disponível em:
<http://www.sciencedirect.com/science/article/pii/S1877042813036859>

Moghaddam, S. e Ester, M. *Aqa: aspect-based opinion question answering. Data Mining Workshops* (ICDMW), IEEE 11th *International Conference on*. IEEE, 2011, pp. 89–96.

Oliveira Júnior, R. L., Esmin, A. A., Coelho, T. A., Araújo, D. L, Silva, L. A., Giroto, R. Uma Ferramenta de Monitoramento Automático de Mensagens de Fóruns em Ambientes Virtuais de Aprendizagem. Anais do XXII SBIE – XV WIIE, 2011.

O'Reilly, T. *What is web 2.0. design patterns and business models for the next generation of software.*
http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-isweb-20.html, September 2005, *[online]*. Disponível em:
http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-isweb-20.html

Pereira, A. N. Uma técnica de zoneamento para indexação de documentos em sistemas de recuperação de informação. Dissertação de mestrado. Instituto Federal de Educação, Ciência e Tecnologia do Ceará, 2010.

Ravi, S. e Kim, J. *Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers. na: AI IN EDUCATION CONFERENCE (AIED), 2007*.

Salton, G. e Buckley, C. *Term-weighting approaches in automatic text retrieval. Journal information processing and management: an international Journal*, vol. 24, cap. 5, 1988, pp. 513-523.

Sebastiani, F. *Machine learning in automated text categorization. ACM Computing Surveys*, Vol. 34, No. 1, Março 2002, pp. 1–47. Disponível em : <
http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf >

Shi, L., Sun, B., Kong, L., e Zhang, Y. *Web forum sentiment analysis based on topics," in Computer and Information Technology*, 2009. CIT'09. *Ninth IEEE International Conference on*, vol. 2. IEEE, 2009, pp. 148–153.

SPSS. CRISP-DM 1.0 *Step-by-step data mining guide,* 2000.

Rolim, V. B., Cordeiro, F. R., Ferreira, R. Reconhecimento de Padrões Aplicados a Comentários de Fóruns Educacionais. Encontro nacional da inteligência artificial e computacional, 2014.

Roncero, V. G. Classificação semi-supervisionada de textos em ambientes distribuídos. Ph.D. dissertation, Universidade Federal do Rio de Janeiro, 2010.

Rutten, A. L. B. *Bayes' theorem: scientific assessment of experience. International jornal of high dilution research*, vol. 9, no. 32, 2010.

Thorsten, J. A *Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Carnegie-mellon univ pittsburgh pa dept of computer science*, 1996.

Weiss, S. M. e Kulikowski, C. A. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann Publishers Inc., São Francisco, CA, USA, 1991.

Zhang, W., Yoshida, T., Tang, X. *A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications*, v. 38, n. 3, p. 2758-2765, 2011.

ANEXO I – CONTEÚDO PROGRAMÁTICO 1

CONTEÚDO PROGRAMÁTICO DE ALGORITMOS E ESTRUTURA DE DADOS I E II

3. CONTEÚDO PROGRAMÁTICO:

3.1 Módulo I - O computador e a resolução de problemas
- História do computador
- As primeiras ferramentas de cálculo
- O computador eletrônico
- Modelo lógico do computador
- Bits e Bytes
- O hardware
- O software
- A resolução de problemas usando o computador
- Problema
- Algoritmo
- Implementação

3.2 Módulo II - Elementos básicos para elaboração dos algoritmos
- Dados, variáveis e comandos básicos
- Tipos de dados
- As variáveis e a atribuição de valores
- Entrada e saída de dados
- Expressões
- Expressões aritméticas
- Expressões lógicas;
- Expressões literais

3.3 Módulo III - Estruturas de controle
- Estruturas de seleção
- Formato "se - então"
- Formato "se - então - senão"
- Formato encadeado
- Estruturas de repetição
- A estrutura "para"
- A estrutura "enquanto"

3.4 Módulo IV - Estruturas de dados
- Estruturas homogêneas básicas
- Vetores
- Matrizes
- Estruturas homogêneas especiais
- Cadeias de Caracteres
- Conjuntos.

3.5 Módulo V - Modularização de algoritmos
- Subalgoritmos
- Definição - parâmetros, retorno
- Variáveis locais e variáveis globais
- Recursividade
- Definição e aplicações

3.6 Módulo VI – Tipos abstratos de dados (TAD)
- TAD e o Paradigma Imperativo
- Atributos e interface - Registro, Vetor de registros;
- Experimentação;
- TAD e o Paradigma Orientado a Objetos
- Atributos e interface - Classes e objetos;
- Experimentação - Classes predefinidas da linguagem Python;

3.7 Módulo VII - Estruturas de dados lineares
- Listas
- Lista seqüencial e lista encadeada, Lista estática e lista dinâmica;
- TAD lista;
- Pilhas
- TAD Pilha;
- Filas
- TAD Fila;

3.8 Módulo VIII - Estruturas de dados não-lineares
- Árvores
- Propriedades, Caminhamentos;
- TAD Árvore;
- Grafos
- Terminologia, Representação de grafos
- TAD Grafo

ANEXO II – CONTEÚDO PROGRAMÁTICO 2

CONTEÚDO PROGRAMÁTICO DE MATEMÁTICA DISCRETA

1. CONTEÚDO PROGRAMÁTICO:

1.1 Noções de Lógica e Técnicas de Demonstração
- Lógica proposicional
- Proposições e Operadores Lógicos
- Tabela-Verdade
- Equivalências Lógicas
- Conseqüências Lógicas
- Recursão
- Noções de Lógica de Predicados
- Métodos de Prova: Prova Direta, Por Contradição e Por Indução

1.2 Teoria dos Conjuntos
- Definição
- Pertinência, Continência e Igualdade
- Operações entre conjuntos
- Demonstração de Propriedades

1.3 Relações, Funções e Seqüências.
- Relações binárias e n-árias
- Relações reflexivas, simétricas, anti-simétricas e transitivas
- Definição das Funções.
- Funções injetivas, sobrejetivas e bijetivas.
- Seqüências e somatórios.
- Noções de cardinalidade de conjuntos infinitos.

1.4 Introdução a Teoria dos Números
- Relação de Divisibilidade
- Algoritmo da Divisão
- MDC
- Números Primos
- Noções de Aritmética Modular
- Fatorial

ANEXO III – ARTIGO PUBLICADO

Reconhecimento de Padrões Aplicados a Comentários de Fóruns Educacionais

Vitor B Rolim*, Filipe Rolim Cordeiro*†, Rafael Ferreira*†
*Departamento de Estatistica e Informática - Universidade Federal Rural de Pernambuco, Brasil
†Centro de Informática - Universidade Federal de Pernambuco, Brasil

***Resumo*—O uso de plataformas educacionais online e fóruns de discussão requer muitas vezes a necessidade de acompanhamento de milhares de usuários. Para esse acompanhamento ser eficiente, é necessária a distinção entre postagens de dúvidas e respostas em tópicos criados nesses fóruns. Neste trabalho é proposta uma solução para identificação automática de três tipos de postagens em fóruns educacionais: dúvidas, respostas e neutras. Para isto, foram aplicadas e analisadas três técnicas de reconhecimento de padrões, entre elas redes bayesianas, árvore de decisão e rede neural. As técnicas foram analisadas observando-se as métricas de precisão, cobertura e *F-measure*. Resultados mostram que a rede neural conseguiu obter um bom índice de classificação, com valor de *f-measure* de 0.84, 0.7 e 0.81 para as postagens das classes dúvida, neutra e resposta, respectivamente. Por fim, a solução proposta pode ainda servir como entrada para sistemas de recomendação de conteúdo em fóruns educacionais.**

I. Introdução

Com o crescente uso da tecnologia como ferramenta de apoio educacional, o uso de Ambientes de Virtuais de Aprendizagem (AVA) [1] tem aumentado cada vez mais. Estes ambientes disponibilizam várias ferramentas para melhorar a iteração entre professores e alunos, onde alguns exemplos são: fórum, blog, wiki, redes sociais, entre outros. As ferramentas citadas são conhecidas como ferramentas sociais ou ferramentas da Web 2.0 [2]. Elas possuem um grande potencial para gerar conteúdo, o que pode ser usado para auxiliar no processo ensino aprendizagem. Porém, é importante que os AVAs ofereçam meios de acompanhamento direto e indireto para garantir o aprendizado do aluno [3].

O acompanhamento direto é aquele realizado sob a supervisão do professor ou tutor. Para isso, o curso tem um plano de ensino e cronograma de atividades que são acompanhados de perto pelos professores e tutores. Então todo material e discussões disponibilizados no AVA é verificado manualmente para que as dúvidas dos alunos sejam resolvidas e seus progressos computados. Contudo, devido à grande quantidade de alunos normalmente inscritos em cursos com plataformas online, torna-se difícil realizar o acompanhamento direcionado para cada aluno.

Para amenizar essa situação, é necessário também realizar o acompanhamento indireto, que é o acompanhamento sem a participação direta do professor. Para isso, é importante ter um sistema automatizado que possa identificar o tipo de dúvida que um aluno possui e direcionar um conteúdo focado na dúvida do aluno [4], [5]. Existem sistemas que lidam diretamente com blogs [6], fóruns [7], wikis [8].

Dentre essas ferramentas o fórum tem uma característica importante. É nele que os alunos postam dúvidas e possíveis respostas para questões levantas pelo professor. Cheng *et al.* [9] realizou um estudo que mostra a efetividade da ferramenta de fórum para melhorar a performance de estudantes em um AVA. Em um acompanhamento direto as postagens de dúvidas devem receber uma solução. Por outro lado, uma resposta pode ser usada para perceber o progresso do aluno, com isso o professor pode pontuá-lo ou pode utilizar o aluno como propagador do assunto entre os colegas [10].

Para realizar o acompanhamento indireto de um fórum educacional de forma eficiente é importante a utilização de sistemas automáticos. Por exemplo, um sistema para recomendar automaticamente materiais direcionados para as dúvidas de cada aluno seria bastante útil. Para isso, o primeiro passo é realizar a identificação do tipo de postagem feita no fórum educacional. Uma vez identificado o tipo de postagem, pode-se verificar se a postagem foi atendida, em caso de dúvida, e direcionar conteúdos voltados para cada tipo de usuário.

Tendo em vista o problema apresentado, este trabalho propõe uma solução computacional para classificação de postagens de alunos em fóruns educacionais. O sistema proposto realiza a classificação de postagens em três categorias: dúvida, postagem neutra e resposta. Essa etapa é importante para no futuro ser realizada a indicação de conteúdos relacionados com postagem de dúvida em questão.

Para avaliar o sistema proposto, um dataset contendo postagens de fóruns educacionais foi criado e foram testadas várias técnicas de reconhecimento de padrões, sendo que as técnicas de rede bayesiana, árvore de decisão e a rede neural foram as que alcançaram melhores resultados chegando a alcançar um valor de métrica *f-measure* de 0,84 em relação a classe dúvida.

O resto deste artigo está dividido como segue: A Seção II descreve os principais trabalhos relacionados à proposta. As técnicas utilizadas para construção do sistema são detalhadas na Seção III. A seção IV apresenta a metodologia seguida nos experimentos. Na seção V os resultados alcançados pelo sistema são apresentados. Por fim, na Seção VI são apresentados as conclusões e os trabalhos futuros.

II. Trabalhos Relacionados

Esta seção é dividida em duas partes. A primeira descreve alguns trabalhos de análise de sentimento em fórum que foram usados como base para a construção do sistema, e a segunda descreve aplicações de mineração de dados em fóruns educacionais.

A. Análise de Sentimento em Fóruns

A aplicação de análise de sentimento em fóruns vem sendo tema de vários trabalhos [7], [11], [12], [13]. Esta seção descreve os trabalhos que foram usados como base para a nossa proposta.

Shi *et al*. [12] faz análise de sentimento baseado em tópicos de fóruns para agrupar postagens de temas relacionados. O principal objetivo deste trabalho é distinguir opiniões diferentes para um tema usando os conteúdos de fóruns. Para isso os autores transformam as postagens do fórum em uma lista de palavra e usa a frequência das mesmas para realizar o agrupamento.

Seguindo a mesma linha do trabalho citado acima, Li e Wu [11] usam técnicas de agrupamento e classificação de texto para identificar fóruns que estão sendo muito acessados e para agrupar suas postagens por tópicos. Mais uma vez os atributos usados para aplicar as técnicas de mineração de texto é a frequência das palavras nos fóruns.

Resende *et al*. [7] mostra a aplicações de técnicas de pré-processamento e mineração de dados em uma coleção textual. Neste trabalho, ele descreve sobre o uso dos algoritmos que podem ser utilizados para a mineração de texto e como se aplica a mineração de texto para diferentes problemas reais.

Moghaddam e Ester [13] revelam aspectos interessantes sobre análise de sentimento aplicada para sistemas de perguntas e respostas e analise de opinião. Eles mostram como fazer a mineração de opinião, e a partir dessa mineração tirar conclusões sobre as respostas dos tópicos criados. Para isso o sistema utiliza técnicas de mineração de texto para classificar se a postagem contém uma opinião positiva ou negativa.

O nosso trabalho utiliza a ideia de usar uma lista de palavras e suas frequências como atributos para realizar a classificação ([12], [11]) utilizando técnicas de mineração de texto utilizadas em trabalhos anteriores [7], [13].

B. Aplicação de Mineração de Texto em Fóruns Educacionais

A aplicação de mineração de texto em ferramentas de ambientes educacionais vêm crescendo consideravelmente. No contexto de fóruns educacionais podemos listar os trabalhos que seguem.

Aurélio *et al*. [14] utiliza técnicas de sumarização automática para ajudar o professor a acompanhar o grande número de postagens nos fóruns em AVAs. Além disto, este trabalho mostra um estudo sobre a melhoria de se utilizar dispositivos móveis para receber informação da plataforma educacional.

Dringus e Ellis [15] propõem um sistema que identifica o nível de participação dos alunos nos fóruns para ajudar o professor a avaliá-los. O trabalho propõe uma lista de parâmetros quantitativos, por exemplo, quantidade de recursos compartilhados, e qualitativos, como presença no fórum. Este trabalho também identifica os tópicos das postagens para facilitar o acesso do professor às discussões de temas específicos.

Fu-Ren Lin *et al*. [16] propõe um sistema de classificação de gênero das postagens dos fóruns de discussão em ambientes educacionais. Utilizando a frequência das palavras como características para o sistema, e aplicando árvore de decisão para classificação. Os gêneros que este sistema tenta identificar são: anúncios, perguntas, interpretação, conflito, afirmação, e outros.

Existem duas diferenças fundamentais deste trabalho para o nosso sistema: (i) nós classificamos as postagens em apenas 3 categorias (dúvida, resposta e comentário neutro); (ii) nós fizemos um comparativo com diferentes técnicas de classificação para encontrar a que obteve melhor resultado.

III. Técnicas de Classificação

A seguir são descritas as técnicas de aprendizagem de máquina utilizadas nesse trabalho para realizar classificação das postagens de fóruns educacionais.

A. Rede Bayesiana

A técnica Rede Bayesiana [17] é uma técnica probabilística baseada no Teorema de Bayes [18] e é uma técnica bastante utilizada em aprendizagem de máquina e reconhecimento de padrões. Esse classificador calcula a probabilidade que uma amostra desconhecida pertença a cada uma das classes possíveis, predizendo a classe mais provável. Para isto, o classificador baseado em rede bayesiana calcula uma distribuição geradora para cada classe do problema através da análise das relações entre as característica envolvidas e as classes de cada instância.

Após treinar um classificador Bayesiano, dado uma amostra ele decide pela classe com maior probabilidade, representado por $P(w_i/x)$. Essa probabilidade é calculada pela equação 1.

$$P(w_i/x) = \frac{P(w_i)\rho(x/w_i)}{\rho(x)}, \tag{1}$$

$$\rho(x) = \sum_{i=1}^{c} P(w_i)\rho(x/w_i), \tag{2}$$

onde $\rho(x)$ é a função de densidade de probabilidade das classes e $\rho(x/w_i)$ é a função de probabilidade de cada classe w_i. O classificador bayesiano realiza uma classificação estatística, sendo completamente baseado em probabilidades.

B. Árvore de Decisão

Árvore de decisão é uma técnica de aprendizado de máquina que utiliza uma estrutura de árvore para avaliar os atributos de uma entrada e retorna uma predição baseada nos valores desses atributos. Essa árvore é estruturada através de vários nós, onde cada nó corresponde a um teste do valor de uma característica do dado de entrada. Os nós da árvore são ligados por ramos, os quais identificam os possíveis valores do teste realizado em cada nó. Por fim, cada nó da folha da árvore representa um valor de retorno. Sendo assim, a árvore de decisão chega a uma decisão através da realização de vários testes. Para isso, o algoritmo é iniciado na raíz da árvore e a percorre realizando testes sobre as características, que correspondem aos nós, do dado de entrada até chegar na folha da árvore. Ao chegar na folha da árvore é retornado como resultado a classificação.

A Figura 1 mostra a estrutura de uma árvore de decisão binária. Em cada nó da árvore é realizado um teste baseado nos valores dos atributos da amostra e o algoritmo percorre a árvore até chegar na folha, onde é retornado uma resposta de classificação. A árvore de decisão utilizada no trabalho, no entanto, não é binária, podendo ter mais de uma saída para cada nó da árvore. Os testes realizados em cada no, para o trabalho proposto, envolve os valores das características selecionadas nos dados de entradas. A resposta árvore, no caso proposto, é a classe da postagem de entrada no classificador.

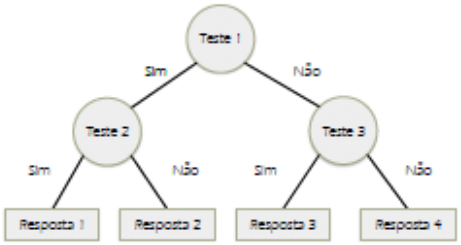


Figura 1. Estrutura de uma árvore de decisão

Existem algumas variações de implementação de árvores de decisão. Nesse trabalho foi utilizada a árvore de decisão J48 [19].

C. Rede Neural

A rede neural é um modelo computacional inspirado nas ligações entre neurônios do cérebro humano. Esse modelo é composto por um conjunto de neurônios, ou nós, que são interligados entre si, formando uma rede. Esses neurônios são divididos em camadas e são conectados por meio de ligações. Cada ligação possui um peso, os são ajustados baseados nas características dos dados de entrada. A Figura 2 mostra a estrutura de uma rede neural.

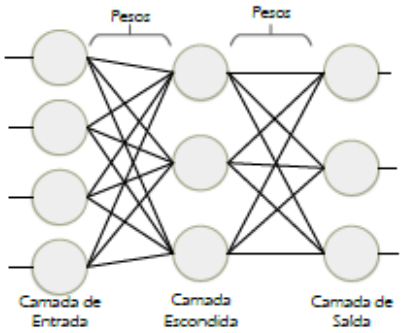


Figura 2. Estrutura de uma rede neural

A partir dos dados de entrada e da força de ligação entre os nós, a rede realiza a classificação dos dados entrada. No trabalho proposto são utilizados 4 neurônios na camada de entrada, onde para cada nó é atribuído uma característica da postagem a ser classificada. Como são extraídas quatro características para cada postagem, que serão descritas na próxima seção, a camada de entrada é composta por 4 neurônios. A camada de saída apresenta três neurônios, onde cada neurônio corresponde a uma classe: pergunta, dúvida ou resposta. O neurônio de saída que obtiver o maior valor final indicará a classe da amostra que está sendo analisada. Nesse trabalho é utilizada a rede MLP (*Multilayer Perceptron*) [20], que é um modelo clássico de rede neural e bastante validado na literatura.

As técnicas de classificação apresentadas são compostas por duas fases: treinamento e testes. Na frase de treinamento é selecionado um conjunto de amostras da base de dados e submetido para cada técnica. Sabendo-se o resultado esperado dessas amostras cada técnica ajusta seus parâmetros automaticamente. Posteriormente, após a fase de treinamento, são feitos os testes outro conjunto dados da base para verificar a precisão das técnicas.

IV. METODOLOGIA

O fluxo deste projeto foi desenvolvido de acordo com a Figura 3.

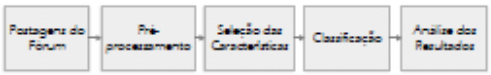


Figura 3. Fluxo de Projeto

Uma vez obtida a base de dados, é realizado o pré-processamento dos dados, onde são removidas as *stopwords*, que são os termos que nada acrescentam à representatividade da coleção de dados, ou que sozinhas nada significam. Exemplos de *stopwords* são palavras como artigos, pronomes e advérbios. O conjunto de *stopwords* é a *stoplist*. Essa eliminação de *stopwords* reduz significativamente a quantidade de termos, diminuindo o custo computacional das próximas etapas [7]. Após a remoção das *stopwords*, é realizada a remoção de termos que tenham uma baixa frequência absoluta em relação à base de dados. Essa remoção é feita porque os termos com baixa frequência normalmente tem pouca influência na identificação das classes definidas. Ainda na etapa de pré-processamento é realizado um ranqueamento de palavras mais frequentes em cada classe: dúvida, resposta e neutra. Esse ranqueamento é importante para saber quais palavras tem mais influência na classificação final de cada classe.

Após essa etapa, é realizada a seleção das características que serão os atributos para as técnicas de classificação. Entre as características utilizadas estão: frequência das palavras da classe Dúvida, frequência das palavras da classe Neutra, frequência das palavras da classe Resposta e número de interrogações. A seleção dessas características é feita contando quantas palavras de cada classe aparecem no texto, para cada postagem. Uma vez obtidos esses atributos, eles são disponibilizados como entrada para as técnicas de classificação. Na etapa de classificação há uma fase de treinamento, onde são ajustados os parâmetros de cada técnica, e uma fase de testes, onde é avaliada a precisão do classifcador. Por fim, é feita a análise dos resultados obtidos por cada um dos classificadores.

A. Seleção das Características

Para a classificação das postagens selecionamos as seguintes características:

o Frequência das palavras da classe Dúvida;
o Frequência das palavras da classe Neutra;
o Frequência das palavras da classe Resposta;
o Número de interrogações na postagem.

A frequência de palavras de cada classe foi utilizada porque existe um conjunto de palavras que aparece com mais frequência para cada tipo de postagem. Sendo assim, é possível atribuir para cada classe um conjunto de palavras que tem maior influência na definição de cada classe de postagem. A Tabela I ilustra o conjunto das cinco palavras mais frequentes em cada classe de postagem.

Tabela I. CONJUNTO DAS CINCO PALAVRAS MAIS FREQUENTES PARA CADA CLASSE DE POSTAGEM

	Duvida	Neutro	Resposta
1	não	semana	direito
2	dúvida	não	não
3	você	disciplina	ser
4	alguém	sobre	forma
5	questão	aqui	você

Conforme mostra Tabela I, a as palavras *dúvida*, *alguém* e *questão* estão entre as palavras mais frequentes da classe dúvida, o que normalmente é encontrado em postagens de alunos com dúvida, em geral. O ranqueamento de palavras mais frequentes para cada classe é feito para mais de 100 palavras por classe, no entanto na Tabela I são mostradas apenas as 5 palavras mais frequentes. Existem palavras que são bastante frequentes para todas as classes, como a palavra *você* e *não*. Essa repetição de palavras não atrapalha o classificador, pois o que conta é a quantidade de palavras diferentes de cada classe para cada postagem.

O uso da característica de número de interrogações por postagem também é um parâmetro interessante, pois o caractere de interrogação está bastante frequente em postagens da classe de dúvida. Embora seja possível encontrar a interrogação em postagens de outras classes, a combinação de todas as características selecionadas fornece um conjunto de parâmetros adequado para as técnicas de classificação.

Outras tipos de características também podem ser aplicados, utilizando os mesmos classificadores. Melhorias no desempenho das técnicas podem ser obtidas através de características que representem melhor cada tipo de postagem.

B. Ambiente Experimental

Para essa pesquisa foi utilizada a ferramenta Weka (*Waikato Environment for Knowledge Analysis*) [21][22], versão 3.6.10. O *Weka* é uma ferramenta que disponibiliza uma coleção de algoritmos de aprendizado de máquina para utilização na mineração de dados. Nele estão implementadas as técnicas utilizadas no trabalho, assim como outras técnicas de aprendizagem de máquina.

Conforme mencionado anteriormetne, todas as técnicas implementadas utilizam uma etapa de treinamento e validação. No processo de classificação dos dados aplicamos o método de *cross-validation* [23], com o número de *folds* igual a 10. Esse método divide os dados em vários grupos, dividindo-os entre treinamento e testes.

Neste trabalho utilizou-se uma base de dados construída a partir das postagens de um fórum educacional acadêmico. A base de dados é distribuída da seguinte forma: postagens que são consideradas como dúvidas; postagens que são consideradas como neutras; postagens que são consideradas como respostas. A base utilizada consiste de 490 postagens, as quais se encontram distribuídas de acordo com a proporção mostrada na Figura 4.

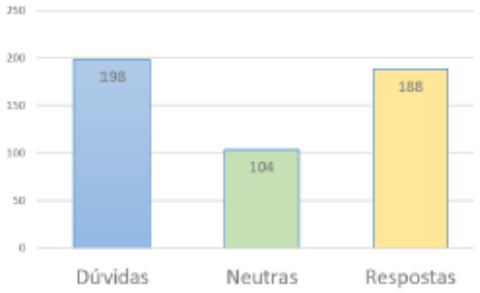


Figura 4. Número de amostras para cada classe

C. Métricas

Neste trabalho foram utilizada as seguintes métricas de análise: Precisão, Cobertura e *F-Measure*.

Segundo Friedman [5], a precisão avalia o quanto o sistema acerta. A métrica de precisão aplicada no trabalho é baseada na quantidade de postagens classificadas corretamente. A precisão é calculada pela Equação 3.

$$Precisao = \frac{tp}{tp + fp}, \tag{3}$$

onde *tp* representa o número de verdadeiros positivos, enquanto *fp* é o número de falsos positivos.

A métrica de Cobertura [24] avalia o quanto a base de dados possui de informação descartável em seu resultado. Podemos observar sua equação abaixo:

$$Cobertura = \frac{tp}{tp + fn}, \tag{4}$$

onde *tp* representa o número de verdadeiros positivos e *fn* é o número de falsos negativo.

A métrica *F-Measure* pode ser interpretada como uma média ponderada da precisão e da cobertura, onde o F tem o seu melhor valor em 1 e o pior em 0, conforme pode ser notado na equação 5.

$$F = 2 * \frac{precisao \cdot cobertura}{precisao + cobertura}, \tag{5}$$

V. RESULTADOS

Esta seção analisa os resultados de classificação das postagens de fórum da base de dados, utilizando os classificadores Naive Bayes, Árvore de Decisão e Redes Neurais.

Aplicando-se o Rede Bayesiana, foi obtida a matriz de confusão apresentada na Tabela II. Essa matriz mostra a relação do número de elementos da classe classe real, representado nas colunas e o número de elementos da classe atribuída pelo classificador, representado em cada linha. Os valores das células dentro de diagonal mostram as classificações feitas corretamente, enquanto as demais células mostram as classificações incorretas. Observando-se a matriz obtida pelo classificação do Rede Bayesiana, é possível observar que a maioria das postagens da classe dúvida foram classificadas corretamente, assim como as postagens da classe resposta. No entanto, para as postagens neutras houve um maior número de classificações erradas de postagens da classe neutra como sendo da classe resposta.

Tabela II. MATRIZ DE CONFUSÃO DA CLASSIFICAÇÃO DA REDE BAYESIANA

Dúvida	Neutro	Resposta	Classificado como
152	37	9	Dúvida
17	74	13	Neutro
22	85	81	Resposta

As métricas de análise foram aplicadas ao classificador de rede bayesiana, cujos resultados são apresentados na Tabela III.

Tabela III. RESULTADOS DAS MÉTRICAS PARA A REDE BAYESIANA

	Dúvida	Neutro	Resposta
Precisão	0.796	0.378	0.786
Cobertura	0.768	0.712	0.431
F-measure	0.781	0.493	0.557

Como pode ser observado na Tabela III, a Rede Bayesiana apresenta um bom desempenho para a métrica de precisão nas classificações de dúvida e resposta. Isso mostra que grande parte da classificação da técnica para essas classes estão corretas. No entanto, para a classe neutro o classicador não obteve uma boa precisão. A métrica de cobertura também indicou bons resultados para a classe de dúvidas. Por fim, a métrica F-measure indica que a técnica Rede Bayesiana consegue classificar corretamente as postagens da classe dúvida, enquanto as classes Neutro e Resposta ela possui maior dificuldade de classificação.

A mesma análise foi aplicada ao algoritmo J48, onde o resultado da matriz de confusão é apresentado na Tabela IV.

Tabela IV. MATRIZ DE CONFUSÃO DA CLASSIFICAÇÃO A ÁRVORE DE DECISÃO J48

Dúvida	Neutro	Resposta	Classificado como
162	18	18	Dúvida
15	64	25	Neutro
23	26	139	Resposta

A matriz de confusão da Tabela IV mostra que a técnica J48 consegue classificar corretamente a maioria das postagens referentes à cada classe, obtendo resultados quantitativos melhores do que a técnica Rede Bayesiana. O resultado de análise das métricas aplicadas é apresentado na Tabela V.

Tabela V. RESULTADOS DAS MÉTRICAS PARA A ÁRVORE DE DECISÃO J48

	Dúvida	Neutro	Resposta
Precisão	0.81	0.593	0.764
Cobertura	0.818	0.615	0.739
F-measure	0.814	0.604	0.751

Conforme descrito na Tabela V, a técnica J48 apresenta bons índices de precisão, cobertura e F-measure para as classes de dúvida e resposta, conseguindo obter resultados melhores quando comparados à técnica Rede Bayesiana. As classificações para a classe Neutro, no entanto, não tiveram resultados tão expressivos.

Por fim, a mesma análise é realizada para a rede neural MLP. Os resultados de classificação dessa técnica é mostrado na Tabela VI

Tabela VI. MATRIZ DE CONFUSÃO DA CLASSIFICAÇÃO A REDE NEURAL MLP

Dúvida	Neutro	Resposta	Classificado como
168	68	14	Dúvida
13	74	17	Neutro
21	15	152	Resposta

Conforme mostrado na Tabela VI, a técnica MLP consegue obter a classificação correta das amostras, para cada classe de postagem, na maioria dos casos. Os resultados quantitativos da matriz de confusão mostra também que o número de classificações corretas para cada classe, que corresponde à diagonal principal da matriz, é maior do que quando comparado às técnicas Rede Bayesiana e J48.

Os resultados das métricas aplicadas à classificação obtida pela MLP é mostrada na Tabela VII.

Tabela VII. RESULTADOS DAS MÉTRICAS PARA A REDE NEURAL MLP

	Dúvida	Neutro	Resposta
Precisão	0.832	0705	0.831
Cobertura	0.848	0.712	0.809
F-measure	0.84	0.708	0.819

A partir da Tabela VII é possível notar que a classificação da MLP obteve valore de precisão, cobertura e F-measure acima de 0.7, para as 3 classes. Isso indica que o classificador conseguiu identificar corretamente cada classe, na maioria dos casos, obtendo poucos resultados de classificação incorreta. Além disso, o classificador obteve ainda melhor desempenho na identificação das classes de dúvida, obtendo valor de F-measure 0.84, sendo maior do que das técnicas Rede Bayesiana e J48.

Ao analisarmos os resultados, chega-se a conclusão que a rede neural MLP é a mais indicada para a classificação de postagens de fóruns educacionais, dentre as três técnicas analisadas. A conclusão de ter um bom classificador para o problema impacta na possibilidade de identificar automaticamente o tipo de postagem, possibilitando ser integrado à um sistema que ajude o aluno quando houver postagem de dúvida. Isso permite a implantação de um modelo educacional voltado ao auxílio personalizado e automático ao aluno, mesmo em ambientes online onde há um grande número de usuários.

VI. Conclusão

Com a grande disseminação da educação a distância os ambientes virtuais de aprendizagem estão sendo largamente usados. Nestes ambientes as ferramentas da Web 2.0, como fóruns, blogs, wiki, tem ganhado um papel no processo de ensino-aprendizagem. Os fórum tem uma característica importante, nele os alunos postam dúvidas e possíveis respostas para questões levantas pelo professor. Com isso, esta ferramenta produz um conteúdo bastante valioso para o curso.

Contudo, devido à grande quantidade de alunos normalmente inscritos em cursos com plataformas online, torna-se difícil realizar o acompanhamento direcionado para cada aluno de forma. Além disto, o acesso ao conteúdo do fórum também não é utilizado de forma eficiente.

Diante deste problema, este artigo propõe a utilização de técnicas de classificação de texto para identificar o gênero da postagem realizada no fórum. Esse experimento foi conduzido em uma base de 490 postagens de fóruns distribuídas entre três classes: dúvida, resposta ou comentário neutro. Com essa informação o professor pode responder as dúvidas e avaliar a participação dos alunos, criar grupos de estudo, recomendar material complementar para o estudo, entre outros.

Foram utilizadas as técnicas rede bayesiana, J48 e MLP, onde foi possível observar que a rede MLP obteve melhores resultados para classificação de postagens dos fóruns, obtendo valor da métrica*f-measure* de 0,84 em relação a classe dúvida, 0.7 para a classe neutro e 0.81 para a classe resposta.

Como trabalhos futuros pretende-se implementar um sistema de recomendações de conteúdo baseado nos resultados obtidos nesse trabalho, uma vez que foi identificado quais os tipos de postagens. Por fim, pretendemos realizar a integração deste sistema a uma plataforma educacional já existente para avaliar o sistema no contexto pedagógico.

Referências

[1] P. S. Pierre Dillenbourg, Daniel Schneider, "Virtual learning environments," in *Proceedings of the 3rd Hellenic Conference on Information & Communication Technologies in Education*, 2002, pp. 3–18.

[2] T. O'Reilly, "What is web 2.0. design patterns and business models for the next generation of software," http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, September 2005, stand 12.5.2011. [Online]. Available: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[3] H. I. Akyuz and M. Kurt, "Effect of teacher's coaching in online discussion forums on students' perceived self-efficacy for the educational software development," *Procedia - Social and Behavioral Sciences*, vol. 9, no. 0, pp. 633 – 637, 2010, world Conference on Learning, Teaching and Administration Papers. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877042810023141

[4] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[5] S. K. Mohamad and Z. Tasir, "Educational data mining: A review," *Procedia - Social and Behavioral Sciences*, vol. 97, no. 0, pp. 320 – 324, 2013, the 9th International Conference on Cognitive Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877042813036859

[6] R. Ferreira, F. Freitas, P. H. S. Brito, J. Melo, R. Lima, and E. Costa, "Retriblog: An architecture-centered framework for developing blog crawlers," *Expert System with Application*, vol. 40, no. 4, pp. 1177–1195, 2013.

[7] S. O. Rezende, R. M. Marcacini, and M. F. Moura, "O uso da mineração de textos para extração e organização não supervisionada de conhecimento," *Revista de Sistemas de Informação da FSMA*, no. 7, pp. 7–21, 2011.

[8] C. R. Nicholas M. Lloyd, Neil T. Heffernan, "Predicting student engagement in intelligent tutoring systems using teacher expert knowledge," in *Proceedings of 13th International Conference of Artificial Intelligence in Education*, 2007, pp. 40–49.

[9] C. K. Cheng, D. E. Par, L.-M. Collimore, and S. Joordens, "Assessing the effectiveness of a voluntary online discussion forum on improving students course performance," *Computers & Education*, vol. 56, no. 1, pp. 253 – 261, 2011, serious Games. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360131510002198

[10] J. Kim, "Influence of group size on students' participation in online discussion forums," *Computers & Education*, vol. 62, no. 0, pp. 123 – 129, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360131512002539

[11] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decision Support Systems*, vol. 48, no. 2, pp. 354–368, Jan. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2009.09.003

[12] L. Shi, B. Sun, L. Kong, and Y. Zhang, "Web forum sentiment analysis based on topics," in *Computer and Information Technology, 2009. CIT'09. Ninth IEEE International Conference on*, vol. 2. IEEE, 2009, pp. 148–153.

[13] S. Moghaddam and M. Ester, "Aqa: aspect-based opinion question answering," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 89–96.

[14] M. A. Gerosa, D. Filippo, M. Pimentel, H. Fuks, and C. J. Lucena, "Is the unfolding of the group discussion off-pattern? improving coordination support in educational forums using mobile devices," *Computers & Education*, vol. 54, no. 2, pp. 528 – 544, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360131509002449

[15] L. P. Dringus and T. Ellis, "Using data mining as a strategy for assessing asynchronous discussion forums," *Computers & Education*, vol. 45, no. 1, pp. 141 – 160, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360131504000788

[16] F.-R. Lin, L.-S. Hsieh, and F.-T. Chuang, "Discovering genres of online discussion threads via text mining," *Computers & Education*, vol. 52, no. 2, pp. 481–495, 2009.

[17] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[18] A. L. B. Rutten, "Bayes' theorem: scientific assessment of experience; teorema de bayes: uma avaliação cientifica da experiência; teorema de bayes: evaluación científica de la experiencia," *Int. j. high dilution res*, vol. 9, no. 32, 2010.

[19] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," *International Journal*, vol. 3, no. 6, 2013.

[20] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher, and P. Held, "Multi-layer perceptrons," in *Computational Intelligence*. Springer, 2013, pp. 47–81.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[22] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg, "Weka-a machine learning workbench for data mining," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2010, pp. 1269–1277.

[23] S. Arlot, A. Celisse *et al.*, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[24] V. G. Roncero, "Classificação semi-supervisionada de textos em ambientes distribuídos," Ph.D. dissertation, Universidade Federal do Rio de Janeiro, 2010.

ANEXO IV – POSTER APRESENTADO

2015 SECOMP UFRPE — II Edição da Semana da Computação UFRPE
Dois Irmãos – Recife – PE
De 19 a 24 de Outubro de 2015

Análise Comparativa Entre Técnicas De Classificação De Texto Para Identificação De Dúvidas Em Fóruns Educacionais

Rolim, V.B.; Ferreira, R.; Cordeiro, F.R.

Departamento de Estatística e Informática (DEINFO)
Universidade Federal Rural de Pernambuco(UFRPE), Brasil

1. Introdução

Com o crescente uso da tecnologia como ferramenta de apoio educacional, o uso de Ambientes Virtuais de Aprendizagem (AVA) [1] tem aumentado nos últimos anos. Estes ambientes disponibilizam várias ferramentas para melhorar a interação entre professores e alunos, onde alguns exemplos são: fórum [2], blog, *wiki*, redes sociais, entre outros. Estas ferramentas possuem um grande potencial para gerar conteúdo, o que pode ser usado para auxiliar no processo ensino aprendizagem. Porém, é importante que os AVAs ofereçam meios de acompanhamento direto e indireto para garantir o aprendizado do aluno [3]. O acompanhamento direto é aquele realizado sob a supervisão do professor ou tutor.

O acompanhamento direto é aquele realizado sob a supervisão do professor ou tutor. Para isso, o curso tem um plano de ensino e cronograma de atividades que são acompanhados de perto por eles. Então todo material e discussões disponibilizados no AVA é verificado manualmente para que as dúvidas dos alunos sejam resolvidas e seus progressos computados. Contudo, devido à grande quantidade de alunos normalmente inscritos em cursos com plataformas online, torna-se difícil realizar o acompanhamento direcionado para cada aluno.

Para amenizar essa situação é necessário também realizar o acompanhamento indireto, que é o acompanhamento sem a participação direta do professor. Para isso, é importante ter um sistema automatizado que possa identificar o tipo de dúvida que um aluno possui e direcionar um conteúdo focado na dúvida do aluno [4].

Tendo em vista o problema apresentado, este trabalho propõe uma solução computacional para análise de identificação de dúvidas em fóruns educacionais.

2. Metodologia

O fluxo deste projeto foi desenvolvido em quatro etapas: pré-processamento, seleção de características, classificação e análise de resultados.

Pré-Processamento: a partir de postagem do fórum é realizada a etapa de pré-processamento do texto. Para esta etapa, duas técnicas foram empregadas: a remoção de *stopwords* e o *lemmatization* [5].
Seleção das características: após a etapa de pré-processamento, é realizada a seleção das características para as técnicas de classificação. Entre as características utilizadas estão: frequência das palavras de cada classe (dúvida, neutra, resposta) e número de interrogações. A seleção dessas características é feita contando quantas palavras de cada classe aparecem no texto, para cada tipo de postagem. Esses características serão usadas como entrada para as técnicas de classificação.

	Dúvida	Neutro	Resposta
1	não	semana	direito
2	dúvida	não	não
3	você	disciplina	ser
4	alguém	sobre	forma
5	questão	aqui	você

Classificação: na etapa de classificação há uma fase de treinamento, onde são ajustados os parâmetros de cada técnica, e uma fase de testes, onde é avaliada a precisão do classificador. Neste trabalho utilizamos a ferramenta Weka (*Waikato Environment for Knowledge Analysis*) [6].
Analise dos Resultados: No processo de avaliação dos resultados, aplicamos o método de validação cruzada com 10 *k-fold* [7]. Para avaliar os classificadores foram usadas as seguintes métricas: Precisão, Cobertura e *F-Measure* [8].

3. Resultados

Inicialmente foram feitos estudos de caso observando-se o impacto da aplicação de elementos de pré-processamento no resultado final da classificação. Dentre os elementos de pré-processamento analisados estão: remoção de *stopwords*, *lemmatization*, remoção de palavras repetidas e variação no limiar de seleção de palavras mais frequentes. As análises foram realizadas para a rede neural MLP.

A princípio foram aplicadas ao banco de dados apenas a técnica de remoção de stopwords e delimitados 3 limiares de frequência de palavras. Esses limiares são utilizados para selecionar o conjunto palavras mais frequentes para serem incorporados na etapa de seleção de características. As palavras com frequência menor que o limiar definido são descartadas. Na análise realizada foram utilizados os limiares 0, 3 e 5, que representam ausência de corte, frequência maior ou igual a 3 e maior ou igual 5, respectivamente.

Análise com remoção de *stopwords*

Classes	F-Measure		
	limiar em 0	limiar em 5	limiar em 10
Dúvida	0.834	0.862	0.833
Neutra	0.528	0.765	0.682
Resposta	0.727	0.801	0.776
Média	0.729	0.818	0.779

Análise com remoção de *stopwords* e *steaming*

Classes	F-Measure		
	limiar em 0	limiar em 5	limiar em 10
Dúvida	0.837	0.837	0.847
Neutra	0.575	0.739	0.718
Resposta	0.747	0.815	0.808
Média	0.747	0.808	0.805

Análise com remoção de *stowords*, *steaming*, e palavras repetidas

Classes	F-Measure		
	limiar em 0	limiar em 5	limiar em 10
Dúvida	0.833	0.843	0.859
Neutra	0.545	0.716	0.714
Resposta	0.74	0.788	0.788
Média	0.737	0.795	0.801

Análise das métricas entre os classificadores MLP, Naive Bayes e J48

Classes	MLP			Naive Bayes			J48		
	Precisão	Cobert.	Fm	Precisão	Cobert.	Fm	Precisão	Cobert.	Fm
Dúvida	0.848	0.875	0.862	0.869	0.788	0.826	0.826	0.831	0.829
Neutro	0.785	0.747	0.765	0.442	0.867	0.585	0.686	0.578	0.627
Resposta	0.804	0.799	0.801	0.81	0.456	0.584	0.727	0.785	0.755
Média	0.818	0.819	0.818	0.756	0.679	0.683	0.759	0.76	0.758

4. Conclusão

Os fóruns possuem uma característica importante, pois nele os alunos postam dúvidas e possíveis respostas para questões levantas pelo professor. Com isso, esta ferramenta produz um conteúdo bastante valioso para o curso. Contudo, devido à grande quantidade de alunos normalmente inscritos em cursos com plataformas online, torna-se difícil realizar o acompanhamento direcionado para cada aluno de forma personalizada.

Diante deste problema, este artigo propõe a utilização de técnicas de classificação de texto para identificar o gênero da postagem realizada no fórum, podendo ser: dúvida, neutra ou resposta, auxiliando assim o professor.

Foram utilizadas as técnicas rede bayesiana, J48 e MLP, onde foi possível observar que a rede MLP obteve melhores resultados para classificação de postagens dos fóruns, obtendo 81,8% de acerto. Além disto, foram discutidos cenários pedagógicos onde o sistema proposto poderia facilitar o processo ensino aprendizagem.

Como trabalhos futuros pretende-se implementar um sistema de recomendações de conteúdo, para postagens que forem classificadas como dúvida [9]. Por fim, pretendemos realizar a integração deste sistema a uma plataforma educacional já existente para avaliar o sistema no contexto pedagógico.

5. Referências

[1] Dillenbourg, P. and Schneider, D. P. S. (2002). Virtual learning environments. In Proceedings of the 3rd Hellenic Conference on Information & Communication Technologies in Education, pages 3–18.
[2] Freitas, M. A. S. (2009). Avaliação da Aprendizagem em ambientes de formação online: aportes para uma abordagem hermenêutica. PhD thesis, UFBA: Faculdade de Educação.
[3] Akyuz, H. I. and Kurt, M. (2010). Effect of teacher's coaching in online discussion forums on students' perceived self-efficacy for the educational software development. Procedia – Social and Behavioral Sciences, 9(0):633 – 637. World Conference on Learning, Teaching and Administration Papers.
[4] Baker, R. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1):3–17.
[5] Hotho, A., Nurnberger, A., and Paas, G. (2005). A brief survey of text mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 20(1):19–62.
[6] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., and Trigg, L. (2010). Weka-a machine learning workbench for data mining. In Data Mining and Knowledge Discovery Handbook, pages 1269–1277. Springer.
[7] Arlot, S., Celisse, A., et al. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4:40–79.
[8] Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. Machine learning, 29(2-3):131–163.
[9] Rolim, V. B., Ferreira, R., Cordeiro, F. R. (2014). Reconhecimento de Padrões Aplicados a Comentários de Fóruns Educacionais. Encontro nacional de inteligência artificial e computacional.

Acknowledgments:

DEInfo

ANEXO V – PSEUDOCÓDIGO DA ESTAPA DE CLASSIFICAÇÃO

```
Função escreveCaracteristicas(B,D,N,R,classe)
    //B representa a base de dados completa
    //D lista das palavras de dúvidas
    //N lista das palavras neutras
    //R lista das palavras de respostas

    B1 <- preProcessamento(B) //base de dados com o pré-processamento
    D1 <- preProcessamento(D) //lista das palavras de dúvidas com pré-processamento
    N1 <- preProcessamento(N) //lista das palavras neutras com pré-processamento
    R1 <- preProcessamento(R) //lista das palavras de respostas com pré-processamento

    i <- 0

    para cada post em B1 faça
        quantDuvida <- 0
        quantNeutra <- 0
        quantResposta <- 0
        quantInterrogacao <- 0

        para cada palavraD em D1 faça
            para palavraP em post faça
                se(palavraD = palavraP) então
                    quantDuvida <- quantDuvida + 1

        para cada palavraN em N1 faça
            para cada palavraP em post faça
                se(palavraN = palavraP) então
                    quantNeutra <- quantNeutra + 1

        para cada palavraR em R1 faça
            para cada palavraP em post faça
                se(palavraR = palavraP) então
                    quantResposta <- quantResposta + 1

        para cada palavraP em post faça
            se(palavraP = '?') então
                quantInterrogacao <- quantInterrogacao + 1

        caracteristicas[i] <- [quantDuvida,quantNeutra,quantResposta,quantInterrogacao,classe]
        i <- i + 1

    devolva caracteristicas

algoritmoClassificador(caracteristicas,post) //MLP, J48, NaiveBayes
```

ANEXO VI – PSEUDOCÓDIGO DA ETAPA DE EXTRAÇÃO DE ASSUNTO

```
Função identificaAssunto(post)
    ementa <- extraiConteudoProgramatico(planoDaDisciplina)

    post1 <- Post(post) //o pré-processamento ocorre internamente

    valorPeso <- 0

    para cada topico em ementa faça
        tipo <- tipoTopico(topico)

        se(tipo = pai) então
            objetoPai <- Ontoclass(topico) //cria o objeto pai
            objetoPai.buscaTermos()
        senão
            objetoFilho <- Node(topico,objetoPai) //cria o objeto filho
            objetoFilho.buscaTermos() //o pré-processamento ocorre internamente
            objetoPai.insereFilho(objetoFilho)

        se(objetoPai.filhos != vazio) então
            para cada filho em objetoPai.filho faça
                contadorPeso <- 0
                para cada termoP em post1 faça
                    para cada termoF em filho.termos
                        se(termoP = termoF) então
                            contadorPeso <- contadorPeso + termoF.peso
                        se(contadorPeso >= valorPeso) então
                            valorPeso <- contadorPeso
                            assunto <- filho.nome
                            se(contadorPeso>post1.peso) então
                                post1.peso <- contadorPeso
                                post1.insereAssunto(assunto)

        senão
            contadorPeso <- 0
            para cada termoP em post1 faça
                para cada termoOP em objetoPai.termos
                    se(termoP = termoOP) então
                        contadorPeso <- contadorPeso + termoOP.peso
                    se(contadorPeso >= valorPeso) então
                        valorPeso <- contadorPeso
                        assunto <- objetoPai.nome
                        se(contadorPeso>post1.peso) então
                            post1.peso <- contadorPeso
                            post1.insereAssunto(assunto)

    devolva post1.assunto
```