

Identificação Automática de Dúvidas em Fóruns Educacionais

Vitor B. Rolim¹, Rafael Ferreira¹, Evandro Costa²

¹DEINFO– Universidade Federal Rural de Pernambuco (UFRPE)
Caixa Postal s/n – 52171-900 – Recife – PE – Brasil

²Instituto de Computação – Universidade Federal Rural de Alagoas (UFAL)
Caixa Postal s/n – 57072-900 – Maceió – AL – Brasil

{victor.b.rolim,rafaelflmello,ebcosta}@gmail.com

Abstract. *With the increased use of online educational platforms, the amount of information to be processed by teachers and students is growing quickly. The forum generates a large amount of information. In this context, this paper proposes an automatic solution to identify doubts in educational forums. The proposed system achieved an accuracy of 97%. Finally, the work proposes scenarios where the teacher could benefit from the use of the proposed solution.*

Resumo. *Com o aumento do uso de plataformas educacionais online, a quantidade de informação para ser processada por professores e alunos vem crescendo rapidamente. Nesse contexto, a ferramenta de fórum gera muita informação e exige que o professor faça o acompanhamento de um grande número de postagens em pouco tempo. Diante disto, este trabalho propõe uma solução automática para a análise de identificação de dúvidas em fóruns educacionais. O sistema proposto chegou a uma precisão de 97%. Por fim, o trabalho aponta cenários onde o professor pode se beneficiar com o uso da solução proposta.*

1. Introdução

A utilização dos Ambientes Virtuais de Aprendizagem (AVA) [Dillenbourg e Schneider 2002] como ferramenta de apoio tem aumentado devido a intensificação da utilização da tecnologia no meio educacional. Os ambientes virtuais disponibilizam recursos variados para ampliar a interação entre o professor e o aluno, por exemplo: fóruns, blogs, wikis, entre outros.

É importante que os AVAs possibilitem meios para que o acompanhamento direto e o indireto aconteçam, a fim de que o aprendizado do aluno seja otimizado [Akyuz e Kurt 2010]. O acompanhamento direto é o que acontece sob o monitoramento do professor ou tutor, enquanto o acompanhamento indireto é realizado através de algum serviço computacional.

Devido a interação entre os participantes nas ferramentas dos AVAs, em especial o fórum, a gestão desse ambiente pelo professor se torna uma tarefa árdua tendo em vista o grande número de postagens, sendo um empecilho para o acompanhamento direto. Por isso, para que esse problema seja amenizado é importante a utilização de sistemas automáticos [Azevedo *et al.* 2011, Gerosa *et al.* 2003], como: sistemas que identificam a relevância de uma postagem de acordo com o tema dela; sistemas que possam recomendar um material de estudo para a dúvida do aluno [Baker e Yacef 2009].

Tendo em vista o problema apresentado, este trabalho propõe uma solução computacional para análise de identificação de dúvidas em fóruns educacionais. O sistema proposto realiza a classificação de postagens em três categorias: dúvida, postagem neutra e resposta.

Para avaliar o sistema, um *dataset* contendo postagens de fóruns educacionais foi criado e foram testadas várias técnicas de classificação de texto, sendo que as técnicas de rede bayesiana, árvore de decisão e a rede neural artificial foram as que alcançaram melhores resultados chegando a obter uma taxa de acerto de 97%. Mais detalhes sobre o desenvolvimento podem ser encontrados na Seção 4.

2. Conceitualização e Trabalhos relacionados

Esta Seção é dividida em duas partes. A primeira explica o conceito de fórum e relata as suas possibilidades para o contexto de educação, e a segunda a descreve alguns trabalhos relacionados a aplicação de mineração de texto em fóruns educacionais.

2.1. Fóruns Educacionais

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.

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.

Segundo [Freitas 2009] o professor pode usar o fórum em diversos contextos: (i) incentivar a criação de laços entre os alunos a partir da discussão de temas específicos da disciplina, (ii) desenvolver a capacidade de debate crítico acerca de algum tema ou assunto, (iii) dar um rápido feedback sobre dúvidas e comentários, (iv) guiar os estudos dos alunos baseados nas suas postagens e (v) avaliar o aluno.

2.2. Aplicação de Mineração de Texto em Fóruns Educacionais

A aplicação de mineração de texto em ferramentas de ambientes educacionais vem crescendo consideravelmente [Baker *et al.* 2011]. No contexto de fóruns educacionais podemos listar os trabalhos que seguem.

[Gerosa *et al.* 2010] utiliza técnicas de sumarização automática para ajudar o professor a acompanhar o grande número de postagens nos fóruns em AVAs. Além

disto, esse trabalho mostra um estudo sobre a melhoria de se utilizar dispositivos móveis para receber informação da plataforma educacional.

[Dringus e Ellis 2005] propõe 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.

[Lin *et al.* 2009] propõe um sistema de classificação de gênero das postagens dos fóruns de discussão em ambientes educacionais, utilizando frequência das palavras como características para o sistema e aplicando também árvore de decisão para classificação. Os gêneros que esse sistema tenta identificar são: anúncios, perguntas, interpretação, conflito, afirmação e outros.

Existem duas diferenças fundamentais desse trabalho para o nosso sistema: (i) classificação das postagens em diferentes categorias: dúvida, resposta e comentário neutro; (ii) comparação com diferentes técnicas de classificação para encontrar a que obteve melhor resultado.

3. Técnicas de Classificação

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

3.1. Classificador Naive Bayes

O classificador Naive Bayes [Bouckaert 2007] é uma técnica probabilística baseada no Teorema de Bayes 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 características envolvidas e as classes de cada instância.

3.2. Árvore de Decisão

Indução de 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 raiz 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. Existem algumas variações de implementação de árvores de decisão. Neste trabalho foi utilizada a árvore de decisão J48 [Bhargava *et al.* 2013].

3.3. Rede Neural Artificial

A rede neural artificial é um modelo computacional inspirado nas ligações entre neurônios do cérebro humano [Kruse *et al.* 2013]. 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 quais são ajustados baseados nas características dos dados de entrada. A partir dos dados de entrada e da força de ligação entre os nós, a rede realiza a classificação dos dados entrada. Neste trabalho é utilizada a rede MLP (*Multilayer Perceptron*) [Kruse *et al.* 2013], que é um modelo clássico de rede neural artificial e bastante validado na literatura.

4. Metodologia

O fluxo deste projeto foi desenvolvido em cinco etapas: coleta da base de dados, pré-processamento, extração das características, classificação e análise de resultados.

4.1. Coleta da base de dados

Esta etapa do classificador diz respeito a criação das bases de dados utilizadas neste projeto. Durante o processo de desenvolvimento desta fase de classificação, foram utilizadas 2 bases de dados diferentes.

A base de dados 1 (BD1) foi retirada do AVA da Universidade Federal de Alagoas (UFAL), e as postagens foram coletadas de várias disciplinas do curso à distância de bacharelado em Sistemas de Informação. Já a base de dados 2 (BD2), foi retirada também do AVA da UFAL, porém, das disciplinas de Algoritmos e Estruturas de Dados I e II, do mesmo curso.

A base de dados 1 é composta por 490 postagens, sendo: 198 de postagens contendo dúvida; 104 postagens que são comentários neutros; 188 de postagens contendo respostas. Já a base de dados 2 é composta por 600 postagens, com a seguinte distribuição: 200 postagens contendo dúvida; 200 postagens que são comentários neutros; 200 postagens contendo respostas.

4.2. Pré-processamento

A partir da 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 *stemming* [Hotho *et al.* 2005].

As *stopwords* são as palavras que com pouca representatividade para os textos, por exemplo: artigos e pronomes. A palavra *stemming* deriva da palavra inglesa *stem*, que significa “radical”, por sua vez, técnica homônima transforma a palavra na sua forma radical, removendo sufixos.

Além dessas técnicas, também foi realizada a remoção de termos que tenham uma baixa frequência absoluta em relação à base de dados. Foram estabelecidos alguns limites de corte, para que palavras com o valor da frequência abaixo do limite sejam descartadas. Essa remoção é feita porque os termos com baixa frequência normalmente têm pouca influência na identificação das classes definidas.

As técnicas de pré-processamento variaram, 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.

4.3. Extração das Características

Após a etapa de pré-processamento, foi realizada a extração das características que serão os atributos das técnicas de classificação. Foram definidos dois conjuntos de características. Entre as características utilizadas no 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 tipo de postagem.

Assim como no pré-processamento, fizemos testes usando conjuntos de características diferentes, 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 (*Term Frequency – Inverse Document Frequency*) [Salton e Buckley 1988] 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.

O uso da característica de número de interrogações por postagem também é um parâmetro interessante, pois o caractere de interrogação é bastante frequente em postagens da classe de dúvida, embora seja possível encontrar a interrogação em postagens de outras classes.

4.4. Classificação e Métricas de Avaliação

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*) [Frank *et al.* 2010], versão 3.6.13 para realizar a classificação e avaliação dos resultados. O Weka é uma ferramenta que disponibiliza uma coleção de algoritmos de aprendizado de máquina para utilização na mineração de dados.

No processo de avaliação dos resultados, aplicamos o método de validação cruzada com 10 *k-fold* [Arlot *et al.* 2010]. Esse método divide o conjunto de dados em

treinamento e testes, de forma a ter uma melhor estimativa de classificação, independente do conjunto de testes.

Para avaliar os classificadores foram usadas as seguintes métricas: Precisão, Cobertura e *F-Measure* [Friedman *et al.* 1997]. Precisão avalia a quantidade de instâncias que foram classificadas corretamente. Cobertura avalia a porcentagem instâncias de uma determinada classe que não foi classificada como pertencente a essa classe. *F-Measure* pode ser interpretada como uma média harmônica da precisão e da cobertura.

5. Análise dos Resultados

Para avaliarmos qual seria a melhor técnica de classificação para atingir o objetivo apresentado, utilizamos as seguintes técnicas contidas no Weka: Naive Bayes, Árvore de decisão (J48), Rede neural artificial (MultilayerPerceptron).

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 1. A avaliação da classificação será feita usando a métrica *F-measure*, quanto mais próximo a *F-measure* for de 1, melhor o resultado.

Tabela 1 - Média das *F-Measure* das técnicas de classificação

<i>Técnicas de classificação</i>	<i>Média da F-measure</i>	
	<i>BD1</i>	<i>BD2</i>
NaiveBayes	0.526	0.701
J48	0.885	0.882
MultilayerPerceptron	0.962	0.972

Ao analisarmos os resultados, chega-se a conclusão que a rede neural artificial MLP é a mais indicada para a classificação de postagens de fóruns educacionais, dentre as três técnicas analisadas, alcançando os valores de 0.962 e 0.972 para BD1 e BD2 respectivamente. A conclusão de ter um bom classificador para o problema impacta na possibilidade de identificar automaticamente o tipo de postagem do fórum, possibilitando ser integrado a um sistema que ajude o aluno quando houver postagem de dúvida. O resultado confirma a conclusão de Rolim *et al.* [Rolim *et al.* 2014], que aponta o MLP como melhor algoritmo classificador para a 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 possa ser encontrado o melhor cenário possível.

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 *F-measure* da classe Dúvida.

5.1. Usando frequência das palavras

Ao utilizar como características as frequências das palavras foi definido 3 cenários diferentes, são eles:

- Cenário 1: Classificação sem técnicas de pré-processamento;
- Cenário 2: Classificação usando remoção de *stopwords*;
- Cenário 3: Classificação usando remoção de *stopwords* e *stemming*.

Os limites de cortes utilizados variaram entre 0, 5 e 10. Podemos observar os valores da *F-measure* alcançados para cada classe de acordo com a base de dados utilizadas na Tabela 2.

Tabela 2 - F-measure das classes de acordo com a base de dados (Frequência).

		Limite de corte 0			Limite de corte 5			Limite de corte 10		
		Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta
BD1	Cenário 1	0.96	0.935	0.979	0.738	0.708	0.774	0.728	0.623	0.745
	Cenário 2	0.963	0.953	0.966	0.793	0.768	0.801	0.736	0.634	0.71
	Cenário 3	0.892	0.825	0.922	0.779	0.777	0.806	0.77	0.713	0.782
BD2	Cenário 1	0.977	0.961	0.977	0.915	0.844	0.835	0.821	0.797	0.827
	Cenário 2	0.962	0.933	0.955	0.906	0.859	0.872	0.891	0.823	0.834
	Cenário 3	0.872	0.874	0.895	0.879	0.809	0.854	0.842	0.785	0.796

O cenário 1 e o cenário 2 apresentaram os maiores valores da *F-measure* da classe dúvida tanto na BD1 quanto na BD2 quando o limite de corte estava em 0.

5.2. Usando TF-ITF

Assim como na Seção 5.1 foram definidos 3 cenários diferentes:

- Cenário 4: Classificação sem técnicas de pré-processamento;
- Cenário 5: Classificação usando remoção de *stopwords*;
- Cenário 6: Classificação usando remoção de *stopwords* e *stemming*.

O limite de corte, 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. A Tabela 3 mostra os valores da *F-measure* para cada classe nas duas bases de dados.

Tabela 3 - F-measure das classes de acordo com a base de dados (TF-IDF).

		Limite de corte 0			Limite de corte 0.00109			Limite de corte 0.00247		
		Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta	Dúvida	Neutra	Resposta
BD1	Cenário 4	0.832	0.48	0.636	0.854	0.819	0.795	0.834	0.705	0.748
	Cenário 5	0.832	0.879	0.78	0.911	0.898	0.87	0.855	0.763	0.788
	Cenário 6	0.823	0.667	0.711	0.908	0.916	0.87	0.868	0.849	0.823
BD2	Cenário 4	0.938	0.908	0.895	0.885	0.817	0.777	0.846	0.655	0.585
	Cenário 5	0.85	0.549	0.536	0.941	0.899	0.895	0.893	0.822	0.8
	Cenário 6	0.843	0.639	0.611	0.926	0.885	0.88	0.882	0.793	0.801

Ao observarmos os resultados apresentados, notamos 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 1 nos mostra os maiores valores atingidos da *F-Measure* da classe Dúvida para as duas bases de dados utilizadas.

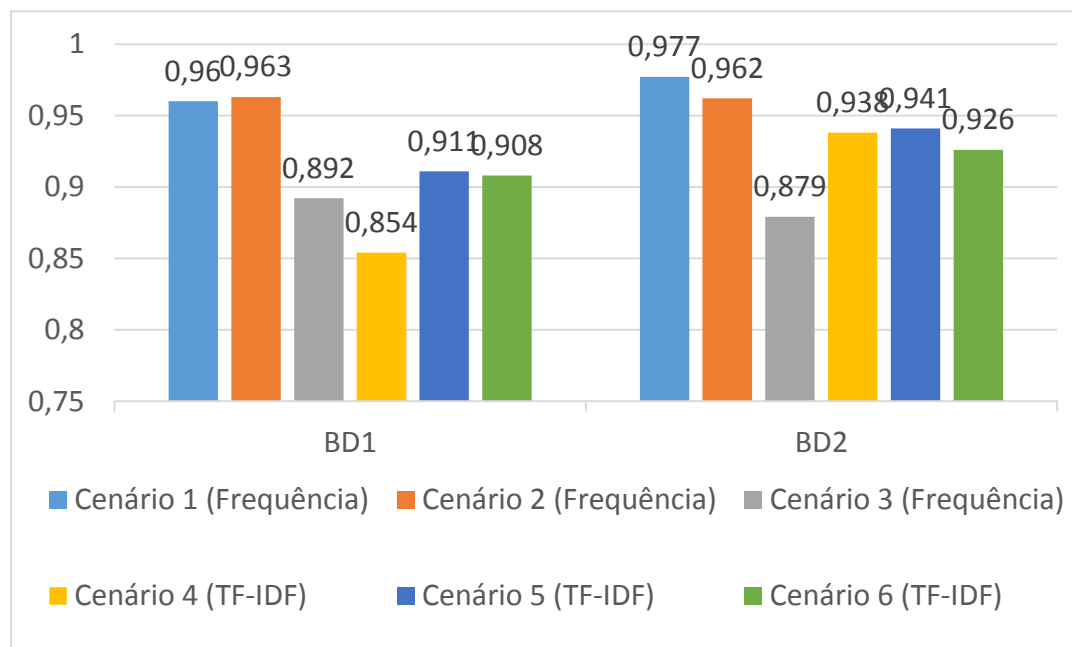


Figura 1 - Maiores valores atingidos da *F-measure* da classe Dúvida.

Podemos observar que o cenário 1 e 2 são os que apresentam os maiores valores da *F-measure* em ambas as bases de dados.

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 1 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 *F-measure*, o que representa uma maior acurácia.

6. Conclusão

Os fóruns possuem uma característica importante, pois neles os alunos postam dúvidas e possíveis respostas para questões levantadas 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 tipo da postagem realizada no fórum. Esse experimento foi conduzido em bases de dados compostas por 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 de rede bayesiana, árvore de decisão e redes neurais, onde foi possível observar que a rede neural artificial (MLP) obteve melhores resultados para classificação de postagens dos fóruns, obtendo 97% de acerto. Ademais, 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 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.

7. Referências

- 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.
- Arlot, S., Celisse, A., et al. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.
- Azevedo, B. F. T., Behar, P. A., Reategui, E. B. (2011). Análise temática das mensagens de discussões online, *Cadernos de Informática – Volume 6 – Número 1*.
- 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.
- Baker, R.S.J.d., Isotani, S., de Carvalho, A. (2011) Mineração de Dados Educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*. 19 (2), 3-13
- Barros, G. e Carvalho, M. A. (2011). As concepções, de interatividade nos ambientes virtuais de aprendizagem. *Tecnologias digitais na educação*, 1(1):209–232.
- Bhargava, N., Sharma, G., Bhargava, R., and Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *International Journal*, 3(6).
- Cheng, C. K., Par, D. E., Collimore, L.-M., and Joordens, S. (2011). Assessing the effectiveness of a voluntary online discussion forum on improving students course performance. *Computers & Education*, 56(1):253 – 261. *Serious Games*.

- Bouckaert, R. R. (2007). Bayesian Network Classifiers in Weka. CIS303: Advanced Forensic Computing.
- 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.
- Dringus, L. P. and Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education*, 45(1):141 – 160.
- 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.
- Freitas, M. A. S. (2009). Avaliação da Aprendizagem em ambientes de formação online: aportes para uma abordagem hermenêutica. PhD thesis, UFBA: Faculdade de Educação.
- Gerosa, M. A., Filippo, D., Pimentel, M., Fuks, H., and Lucena, C. J. (2010). Is the unfolding of the group discussion off-pattern? improving coordination support in educational forums using mobile devices. *Computers & Education*, 54(2):528 – 544.
- 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.
- Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M., and Held, P. (2013). Multilayer perceptrons. In *Computational Intelligence*, pages 47–81. Springer.
- Lin, F.-R., Hsieh, L.-S., and Chuang, F.-T. (2009). Discovering genres of online discussion threads via text mining. *Computers & Education*, 52(2):481–495.
- Rolim, V. B., Ferreira, R., Cordeiro, F. R. (2014). Reconhecimento de Padrões Aplicados a Comentários de Fóruns Educacionais. Encontro nacional da inteligência artificial e computacional.
- 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.