Classificação Automática da Presença Social em Discussões Online Escritas em Português

Jean B. Teixeira[1], Evandro de B. Costa[1], Rafel F. Mello[2], Máverick Ferreira[3], André N. Camara[2]

[1]Instituto de Computação – Universidade Federal de Alagoas (UFAL)

[2]Centro de Informática - Universidade Federal Rural de Pernambuco (UFRPE)

[3]Centro de Informática – Universidade Federal de Pernambuco (UFPE)

{jbt,evandro}@ic.ufal.br, {rafel.mello,andre.camara}@ufrpe.br,

madf@cin.ufpe.br

Abstract. *This work presents a method that allows the automatic classification of messages exchanged in online distance learning forums written in Brazilian Portuguese according to categories (Affective, Interactive and Cohesive) of social presence. To achieve this goal, the adopted method uses a set of 116 resources extracted from text mining and word counting techniques, such as LIWC and Coh-Metrix. The classifier with the best performance presented 0,97 % and 0,95 % for precision and cohen kappa, respectively. This work also provides an analysis of the nature of social presence, looking at the most relevant classification characteristics to distinguish the three categories of social presence.*

Resumo. *Este trabalho apresenta um método que permite a classificação automática das mensagens trocadas em fóruns online de ensino a distância escritas em português brasileiro de acordo com as categorias (Afetiva, Interativa e Coesiva) da presença social. Para atingir esse objetivo, o método proposto faz uso de um conjunto de 116 características extraídas de técnicas de mineração de texto e contagem de palavras como o LIWC e Coh-Metrix. O classificador com melhor desempenho obteve 0,97% e 0,95% para acurácia e cohen kappa, respectivamente. Este trabalho também fornece uma análise da natureza da presença social, observando as características de classificação que foram mais relevantes para distinguir cada uma das três categorias.*

1. Introdução

O Ambiente Virtual de Aprendizagem (AVA) é geralmente utilizado como um mecanismo para facilitar a interação entre professores/tutores e estudantes em cursos de educação *online*. Os AVAs apresentam diversos recursos que proporcionam essa interação, dentre os quais, um dos mais populares é o fórum de discussão [Soares et al. 2016]. Os fóruns de discussão são ferramentas assíncronas que proporcionam uma enorme interatividade entre alunos e professores [Barros 2011], permitindo postagem de dúvidas, cometários sobre o conteúdo da disciplina, postagens de materiais extras, entre outros. Uma pesquisa em aprendizagem *online* e educação a distância aponta que o envolvimento em fóruns de discussão assíncronos acarreta em uma melhora dos resultados acadêmicos

DOI: 10.5753/cbie.sbie.2020.942

[Suhang et al. 2014]. Portanto, analisar as interações entre os alunos e professores torna-se bastante relevante para o processo pedagógico [Garrison et al. 1999].

Na perspectiva de interação mencionada, o presente estudo foca no modelo da Comunidade de Investigação (do inglês *Community of Inquiry (CoI)*. CoI é um dos modelos pedagógicos desenvolvido para apoiar os professores/tutores no desenvolvimento de experiências de aprendizagem *online* moderna [Garrison et al. 1999]. O CoI estabelece três elementos, conhecidos como presenças, que modelam o aprendizado *online* dos alunos, quais sejam: presença social, presença de ensino e presença cognitiva. Dentre elas, a presença social demonstra ser um elemento importante para o sucesso da experiência educacional [Garrison et al. 1999] e também é considerada relevante na observação da forma com que os alunos se lançam nas interações e em sua manutenção nos cursos a distância [Palloff and Pratt 2004].

Dado o contexto descrito anteriormente, este trabalho aborda o problema da identificação automática das categorias de presença social, em mensagens de fóruns educacionais escritas em português brasileiro, propondo uma solução via mineração de texto para desenvolver três classificadores. A principal contribuição é a utilização de características linguísticas, LIWC e Coh-Metrix, que serão detalhadas na Seção 3.4, além de apontar quais características são as mais relevantes para a predição das categorias. Os resultados são analisados sob a perspectiva de aprendizagem colaborativa na área de CSCL (do inglês *Computer-Supported Collaborative Learning*).

2. Fundamentação Teórica e Trabalhos Relacionados

2.1. O Modelo de Comunidade de Investigação (CoI)

O Modelo de Comunidade de Investigação (CoI), é um modelo conceitual proposto por [Garrison et al. 1999] com foco no processo social de construção conjunta e colaborativa do conhecimento em ambientes de comunicação assíncrona baseada em texto. Ele é bastante utilizado para orientar a pesquisa e a prática da aprendizagem *online*, na qual uma comunidade de investigação é constituída por meio de três dimensões, também conhecidas como presenças, essenciais para que haja uma experiência educacional de sucesso. Segundo [Garrison et al. 1999], são elas: (i) A presença social que mede a capacidade de humanizar os relacionamentos entre os participantes de uma discussão; (ii) A presença cognitiva que esta fortemente relacionada ao desenvolvimento dos resultados da aprendizagem; (iii) A presença de ensino que refere-se ao papel do professor antes (isto é, da concepção do curso) e durante o curso.

Este trabalho tem como foco na dimensão da presença social, definida por [Garrison et al. 1999] como a capacidade dos participantes de uma comunidade de investigação de se projetarem social e emocionalmente, como pessoas reais, ou seja, sua personalidade completa, através do meio de comunicação em uso. A presença social, uma das três dimensões do modelo CoI, possui três categorias: (i) Afetiva: categoria que analisa a tradução de emoções reais no texto. Contempla emoções, sentimentos e expressões de humor; (ii) Interativa: tem como foco a interatividade das mensagens trocadas entre os participantes. Seu principal objetivo é melhorar a comunicação aberta entre os alunos; (iii) Coesão de grupo: investiga o senso de união e compromisso de grupo entre os alunos.

2.2. Análises Automáticas do CoI

A Análise de Conteúdo Quantitativo (QCA), é apontada como um método largamente adotado no contexto das três presenças de CoI [Strijbos et al. 2006] com o objetivo de medir/avaliar os processos relacionados a construção do conhecimento nas discussões *online* e fornecer suposições válidas e confiáveis a partir da análise de dados textuais [Bauer 2007]. Em [Garrison et al. 2001], os autores definiram esquemas de codificação para analisar as presenças sociais e cognitivas, os quais têm sido amplamente adotados para a análise de conteúdo manual de CoI.

Em [Kovanovic et al. 2014], os autores utilizaram a codificação manual para avaliar a associação entre presença social e a posição na rede social. As primeiras propostas para automatizar a análise de conteúdo de acordo com os esquemas de codificação do modelo CoI baseavam-se principalmente em recursos tradicionais utilizados em mineração de texto, por exemplo contagem de palavras e frases. Um estudo mais recente utilizou recursos baseados no LIWC e Coh-Metrix para identificar fases da presença social em mensagens escritas em inglês no qual o melhor classificador atingiu 0,95 e 0,88 em acurácia e kappa, respectivamente [Ferreira et al. 2020].

Apesar de existirem estudos para extrair de forma automática as fases do desenvolvimento da presença cognitiva e categorias da presença social em inglês, até o momento não foram encontradas publicações que abordem a classificação automática da presença social em mensagens assíncronas de ambientes *online* de discussão em português.

3. Metodologia

3.1. Questões de Pesquisa

De acordo com o que foi abordado na Seção 2.1, a presença social tem sua importância no CoI pois influencia o desenvolvimento da presença cognitiva nos ambientes virtuais de aprendizagem. Dessa forma, a primeira pergunta de pesquisa deste trabalho é:

Questão de Pesquisa 1 (QP01): *Até que ponto os métodos de mineração de texto podem classificar automaticamente as mensagens de discussão online de acordo com as categorias da presença social?*

Além dessa questão citada acima, pretende-se também disponibilizar informações sobre quais as características que são mais relevantes para classificar cada uma das três categorias. Para isso, foram utilizados alguns parâmetros aplicados por [Kovanović et al. 2016], [Neto et al. 2018] e [Ferreira et al. 2020] com essa mesma finalidade. Então, a segunda questão de pesquisa é:

Questão de Pesquisa 2 (QP02): *Quais características melhor preveem cada categoria da presença social?*

3.2. Descrição do *Corpus*

O *corpus* utilizado foi gerado através de mensagens trocadas em fóruns de discussão de um curso de graduação de Biologia, oferecido totalmente *online* por uma universidade pública brasileira. Foram extraídas 1.500 mensagens produzidas por 215 alunos durante quatro semanas de curso. O objetivo do fórum era de promover discussões sobre um tema proposto pelo professor, em que a participação representava 20% da nota final do

curso. Basicamente as discussões foram do tipo pergunta e resposta, ou seja, os fóruns eram iniciados com uma pergunta pelo professor e os alunos deveriam responder deixando suas contribuições.

Dois codificadores anotaram o conjunto de dados levando em consideração os 12 indicadores da presença social assim como foi feito em [Kovanovic et al. 2014]. Cada mensagem do conjunto de dados para cada indicador recebeu o valor "um" (possui o indicador) ou o valor "zero" (não possui o indicador). Assim como em [Kovanovic et al. 2014], três indicadores (Continuar uma conversa, Expressar apreço/concordância e Vocativos) foram removidos pois continham um grande número de mensagens.

Por fim, como o objetivo neste trabalho foi construir classificadores binários para cada categoria da presença social, as categorias foram compostas pelos indicadores anotados. Para que uma mensagem seja classificada como positiva (1), ela deve ter ao menos um indicador da respectiva categoria anotado com o valor "um". Por exemplo, se uma mensagem continha os indicadores A1 = 0, A2 = 0 e A3= 1 , então ela era considerada positiva para a categoria afetiva.

3.3. Preparação dos Dados de Treinamento e Teste

De acordo com a revisão sistemática da literatura apresentada em [Ferreira-Mello et al. 2019], pode-se constatar que os estudos que se concentram na classificação de texto aplicam algoritmos de aprendizagem de máquina em um conjunto de dados divididos em dois subconjuntos, treinamento e teste, em que o subconjunto de treinamento previamente rotulado é utilizado para gerar um modelo que seja capaz de prever casos futuros, ou seja, prever exemplos da base de teste dos quais os rótulos são desconhecidos. Com base nisso o conjunto de dados foi subdividido em 75% para treinamento e 25% para teste conforme a Tabela 1.

Tabela 1. Distribuição das mensagens entre os grupos de treinamento e teste

Categoria	Grupo	Classe negativa (0)	Classe positiva (1)	Total
Afetiva	Treino	1088 (97%)	37 (3%)	1125
	Teste	367 (98%)	8 (2%)	375
	Total	**1455**	**45**	**1500**
Interativa	Treino	486 (43%	639 (57%)	1125
	Teste	169 (45%	206 (55%)	375
	Total	**655**	**845**	**1500**
Coesiva	Treino	988 (88%)	137 (12%)	1125
	Teste	341 (91%)	34 (9%)	375
	Total	**1329**	**171**	**1500**

3.4. Extração de Características

Além das características tradicionais de mineração de texto neste trabalho utilizou-se as ferramentas linguísticas LIWC e Coh-Metrix para extrair indicativos da presença social nos textos. Para esse trabalho foram utilizadas um total de 116 características extraídas através de ferramentas apresentadas a seguir.

Linguistic Inquiry and Word Count (LIWC): É uma ferramenta desenvolvida por [Pennebaker et al. 2001] com o intuito de fornecer um método eficiente para estudos sobre fatores emocionais, psicológicos, cognitivos entre outros, presentes em trechos de falas verbais e escritas de indivíduos. Em sua versão original em inglês o dicionário possui aproximadamente 6.540 palavras e cada uma associada a uma ou mais categorias dentre as 73 disponíveis. Neste trabalho foi utilizada a versão em português que possui cerca de 127.149 palavras que estão assinaladas a uma ou mais das 64 categorias (social, afetiva, concordância dentre outras) disponíveis [Balage Filho et al. 2013]. Ao relacionar a definição de presença social bem como seus indicadores, hipoteticamente sua utilização pode contribuir para o desenvolvimento dos classificadores capazes de diferenciar de forma correta mensagens com ou sem evidência da presença social. Em [Ferreira et al. 2020], os autores utilizaram a versão em inglês da ferramenta.

Coh-Metrix: Para este trabalho foi utilizada a versão da ferramenta para o português, o Coh-Metrix-PT [1], que possui 48 medidas implementadas de nível léxico, sintático em nível de sintagmas nominais, semântico, e discursivo [Scarton et al. 2010]. Sendo assim, supomos que o índice de coesão e complexidade textual proposto na ferramenta pode relacionar-se com a existência/ausência de indicadores da presença social.

Características de Contexto da Discussão: Com o objetivo de incorporar mais informações de contexto ao conjunto de características neste trabalho, foram incluídas quatro características de contexto utilizadas em [Kovanović et al. 2016]. São elas: (i) Número de respostas, (ii) Profundidade da mensagem, (iii) Similaridade de Cosseno para a mensagem anterior e (iv) Similaridade de Cosseno para a mensagem seguinte.

Frequência de palavras Também foi utilizada uma técnica tradicional de mineração de texto conhecida como *Bag of Words* (BoW), que é uma representação do texto através de uma matriz composta pela quantidade de ocorrência de cada palavra. Ao utilizar essa técnica é comum ocorrer um problema de alta dimensionalidade do vetor devido ao vasto vocabulário presente no texto. Para resolver esse problema e diminuir a dimensionalidade da matriz de ocorrência de palavras foram utilizadas algumas técnicas para limpar o texto removendo URL's, normalizando o texto, removendo *stopwords* que são palavras de pouca importância no texto. Por fim, também foi utilizada uma técnica que visa a redução das palavras aos seus radicais [Orengo and Huyck 2001], por exemplo, as palavras "concordamos" e "concordo" se tornam "concord".

3.5. Pré-processamento dos Dados

Conforme apresentado na Tabela 1, as classes negativas e positivas em todas as categorias apresentam desbalanceamento. Segundo [He and Garcia 2009], lidar com conjuntos de dados que possuem distribuições de classe desequilibradas é um grande desafio pois geralmente as classes majoritárias são priorizadas pelos indutores. Por exemplo, no conjunto de dados utilizados neste trabalho, a categoria afetiva possui 3% de ocorrências da classe positiva e 97% da classe negativa, sugerindo que o classificador pode priorizar a classe negativa. Para resolver esse problema utilizou-se o algoritmo SMOTE que permite a criação de dados artificiais da classe minoritária (*oversampling*) [Chawla et al. 2002].

[1] http://143.107.183.175:22680/

3.6. Seleção e Avaliação do Modelo

Existem diversos algoritmos de aprendizado de máquina para construção de modelos supervisionados. Em [Fernández-Delgado et al. 2014] foi realizada uma análise comparativa utilizando-se de 179 algoritmos de classificação de propósito geral em 121 conjuntos de dados diferentes. Nesse estudo, o algoritmo *Random Forest* foi um dos que apresentaram melhor desempenho. Este trabalho utilizou esse algoritmo não apenas pelo seu ótimo desempenho mas também por ser um algoritmo caixa branca podendo assim identificar quais as características que mais influenciaram na classificação das categorias da presença social. A medida mais utilizada para realizar essa avaliação da importância das características é o *Mean Decrease Gini* (MDG), que explica a separabilidade de uma determinada característica em relação as categorias [Breiman 2001]

Para o algoritmo *Random Forest*, foram estabelecidos dois parâmetros: (i) *n_estimators*: o número de árvores geradas pelo algoritmo; e (ii) *max_features*: o número de características aleatórias selecionadas por cada árvore. Os valores para cada um deles foram baseados na etapa de otimização de parâmetros realizada por [Ferreira et al. 2020] onde constatou-se que os valores de acurácia e kappa se estabilizavam ao utilizar em *max_features* e *n_estimator*, os valores 2.000 e 800 respectivamente.

Nesta etapa, foram utilizados os recursos da biblioteca em *python* denominada *scikit-learn* pois ela possui uma série de funções para serem aplicadas no pré-processamento, treinamento e validação de algoritmos de aprendizado de máquina que supriram as necessidades neste trabalho.

4. Resultados e Discussões

4.1. Modelo de Treinamento e Avaliação - QP01

Como mencionado anteriormente, os valores dos dois parâmetros (*max_features* e *n_estimator*) utilizados no classificador *Random Forest* foram baseados no trabalho de [Ferreira et al. 2020] onde encontrou-se bons valores para eles. Sendo assim foram utilizados o conjunto de dados de treinamento (1.125 mensagens) para gerar um classificador binário de cada categoria, e sua capacidade de generalização foi verificada no conjunto de teste (375 mensagens), os resultados para cada categoria podem ser observados na Tabela 2.

Tabela 2. Resultados dos classificadores por categoria

Categoria	Acurácia	Kappa
Afetiva	0,9786	0,1931
Interativa	0,976	0,9516
Coesiva	0,9786	0,8706

A Tabela 3 mostra a matriz de confusão gerada para cada categoria. É possível notar que as ocorrências mais altas de falsos positivos foram na classe Afetiva com 7 exemplos em um universo de 8 instâncias positivas enquanto as demais categorias tiveram menos de 2% de falsos positivos. Vale ressaltar que a categoria afetiva no conjunto de dados de teste possuía a menor quantidade de instâncias positivas, tornando difícil para o classificador aprender efetivamente como reconhecer as mensagens da categoria.

Tabela 3. Matriz de confusão por categoria

	Afetiva		**Interativa**		**Coesiva**	
	neg*	pos*	neg*	pos*	neg*	pos*
neg*	366	1	167	2	337	4
pos*	7	1	7	199	4	30

*pos = classe postiva e neg = classe negativa

4.2. Análise das Características Importantes - QP02

Este trabalho também analisou as contribuições das diversas características para o desempenho final do classificador. Apesar de terem sido utilizados os mesmos vetores de características para discriminar as classes (positivas e negativas) das três categorias, cada classificador possui um conjunto de variáveis diferentes consideradas como mais importantes. O algoritmo *Random Forest* usa a medida do índice de impureza média de diminuição de gini (MDG) para definir o grau de relevância de uma característica. As Tabelas 4, 5 e 6 apresentam as 15 variáveis mais importantes para o classificador de cada categoria (Afetiva, Interativa e Coesiva).

Para a categoria afetiva o conjunto das variáveis mais importantes apresentadas na Tabela 4 contém dez variáveis de frequência de palavras, quatro LIWC e uma Coh-Metrix. As duas mais importantes foram a palavra "ser" e a variável cm.DESSC, atingindo valores de MDG 12,91 e 9,33 respectivamente.

Tabela 4. Quinze variáveis mais importantes para a categoria afetiva

Variável	**Descrição**	**MDG**
ser	Frequência de palavras	12,91
cm.DESSC	Contagem de frases, número de frases	9,33
liwc.see	Número de palavras que fazem referência a visão	4,47
ter	Frequência de palavras	2,90
morr	Frequência de palavras	2,40
acredt	Frequência de palavras	2,24
brasil	Frequência de palavras	1,94
nov	Frequência de palavras	1,79
def	Frequência de palavras	1,79
vinic	Frequência de palavras	1,71
liwc.friend	Número de palavras que fazem referência a amizade	1,53
profes	Frequência de palavras	1,50
liwc.i	Primeira pessoa do singular	1,42
liwc.hear	Número de palavras que fazem referência a audição	1,29
individu	Frequência de palavras	1,22

Para a categoria interativa, conforme a Tabela 5, temos um conjunto com sete variáveis LIWC, seis frequência de palavras, uma Coh-Metrix e uma do contexto de discussão. A mais importante foi a posição da mensagem dentro da discussão (message.depth) com MDG de 64,88. Também vale destacar que a variável liwc.wd6letters (palavras com mais de 6 letras) também atingiu uma pontuação considerável de 10,58.

Tabela 5. Quinze variáveis mais importantes para a categoria interativa

Variável	Descrição	MDG
message.depth	Posição da mensagem dentro da discussão	64,88
liwc.wd6letters	Palavras com mais de 6 letras	10,58
cm.DESWC	Contagem de palavras, número de palavras	6,94
liwc.words	Quantidade de palavras	3,52
liwc.incl	Número de palavras que fazem referência a inclusão	1,15
liwc.article	Número de artigos	0,9
liwc.preps	Número de preposições	0,56
celul	Frequência de palavras	0,45
liwc.funct	Quantidade de palavras funcionais	0,42
opc	Frequência de palavras	0,35
liwc.cogmech	Número de palavras que fazem referência a cognição	0,33
fer	Frequência de palavras	0,33
dign	Frequência de palavras	0,22
boa	Frequência de palavras	0,20
permit	Frequência de palavras	0,19

Tabela 6. Quinze variáveis mais importantes para a categoria coesiva

Variável	Descrição	MDG
boa	Frequência de palavras	35,99
noit	Frequência de palavras	14,99
ola	Frequência de palavras	8,71
cm.DESPC	Número de parágrafos	4,99
bom	Frequência de palavras	4,00
tarde	Frequência de palavras	3,78
dia	Frequência de palavras	2,29
message.depth	Posição da mensagem dentro da discussão	1,78
obrig	Frequência de palavras	1,70
cm.DESSC	Número de sentenças	1,47
m.DESPL	Número médio de frases em cada parágrafo do texto	1,42
liwc.wd6letters	Palavras com mais de 6 letras	1,08
liwc.swear	Número de palavras que fazem referência a xingamento	0,78
liwc.cogmech	Número de palavras que fazem referência a cognição	0,70
cm.DESSL	Duração da frase, número de palavras, média	0,69

Finalmente, a Tabela 6, aponta as principais variáveis da categoria coesiva, nas quais sete são frequência de palavras,quatro Coh-Metrix, três LIWC, e uma do contexto de discussão. Percebe-se que palavras comumente utilizadas para cumprimentar / saudar obtiveram boas notas (’boa’ - MDG de 35,99, ’noit’ - MDG de 14,99 e ’ola’ - MDG de 8,71). Dessa forma, as variáveis mais importantes apresentadas nas tabelas acima, estão alinhadas com a teoria da presença social proposta por [Garrison et al. 1999], por exemplo, podemos destacar mensagens que possuem: i) palavras que expressam emoções positivas; ii) quantidade de pronomes em segunda pessoa; iii) palavras que expressam concordância; iv) palavras que fazem referência a inclusão; v) palavras que indicam saudações.

5. Conclusão

Este trabalho abordou o problema da identificação automática das categorias de presença social em mensagens de fóruns educacionais, escritas em português. Neste sentido, podemos destacar duas contribuições: a modelagem do problema sob a ótica de aprendizagem de máquina supervisionada, no qual foram propostos três classificadores binários, um para cada categoria da presença social: afetiva, interativa e coesiva. Para isso utilizou-se o algoritmo *Random Forest* e os recursos linguísticos LIWC, Coh-Metrix, contexto da discussão e frequência de palavras (BoW). Com exceção da categorias afetiva, as demais atingiram os valores de *Cohen's kappa* acima de 0,80, que é um acordo entre avaliadores muito bom. A outra contribuição foi a identificação das quinze variáveis mais importantes para o classificador de cada categoria. Importante destacar que com a utilização desses classificadores, os professores/tutores podem identificar o nível da presença social do grupo ou de cada aluno e realizar uma intervenção no decorrer do curso para melhorá-lo e consequentemente auxiliar no processo de ensino-aprendizagem.

Vale ressaltar que a pesquisa apresentou algumas limitações, podemos destacar o problema com o tamanho pequeno da base de dados utilizadas e as categorias desbalanceadas, apesar de que essa situação reflita com as encontradas na literatura, podem afetar o desempenho do classificador. Para trabalhos futuros pretende-se utilizar uma amostra maior de dados composta por diferentes domínios, bem como realizar uma etapa de otimização buscando ajustar os parâmetros utilizados para melhorar os resultados obtidos principalmente na categoria afetiva. Também está sendo desenvolvida uma versão mais completa da ferramenta Coh-Metrix para o português brasileiro.

Referências

Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Barros, Maria das Graças e Carvalho, A. B. G. (2011). As concepções de interatividade nos ambientes virtuais de aprendizagem. *Campina Grande: EDUEPB*.

Bauer, M. W. (2007). Content analysis. an introduction to its methodology–by klaus krippendorff from words to numbers. narrative, data and social science–by roberto franzosi. *The British Journal of Sociology*, 58(2):329–331.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181.

Ferreira, M., Rolim, V., Mello, R. F., Lins, R. D., Chen, G., and Gašević, D. (2020). Towards automatic content analysis of social presence in transcripts of online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 141–150.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Garrison, D. R., Anderson, T., and Archer, W. (1999). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The internet and higher education*, 2(2-3):87–105.

Garrison, D. R., Anderson, T., and Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of distance education*, 15(1):7–23.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Kovanovic, V., Joksimovic, S., Gasevic, D., and Hatala, M. (2014). What is the source of social capital? the association between social network position and social presence in communities of inquiry. In *Proceedings of the Workshops held at Educational Data Mining 2014 co-located with 7th International Conference on Educational Data Mining (EDM 2014)*. Citeseer.

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., and Siemens, G. (2016). Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 15–24.

Neto, V., Rolim, V., Ferreira, R., Kovanović, V., Gašević, D., Lins, R. D., and Lins, R. (2018). Automated analysis of cognitive presence in online discussions written in portuguese. In *European conference on technology enhanced learning*, pages 245–261. Springer.

Orengo, V. M. and Huyck, C. R. (2001). A stemming algorithmm for the portuguese language. In *spire*, volume 8, pages 186–193.

Palloff, R. M. and Pratt, K. (2004). *O aluno virtual-um guia para trabalhar com estudantes on-line*. Penso Editora.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Scarton, C., Gasperin, C., and Aluisio, S. (2010). Revisiting the readability assessment of texts in portuguese. In *Ibero-American Conference on Artificial Intelligence*, pages 306–315. Springer.

Soares, F. B. M., Machado, C. J. R., Diniz, D., and Maciel, A. M. A. (2016). Educational data mining to support distance learning students with difficulties in the portuguese grammar. In *Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE 2016)*, pages 956–965, Brasil.

Strijbos, J.-W., Martens, R. L., Prins, F. J., and Jochems, W. M. (2006). Content analysis: What are they talking about? *Computers & Education*, 46(1):29–48.

Suhang, J., Williams, A., Schenke, K., Warschauer, M., and Odowd, D. (2014). Predicting mooc performance with week 1 behavior. *Educational Data Mining*.