Uma Nova Abordagem para Detecção de Plágio em Ambientes Educacionais

Anderson P. Cavalcanti[1], Rafael Ferreira[1], Máverick A. D. Ferreira[1], Sebastião Neto[2], Guilherme Passero[3], Péricles Miranda[1]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco (UFRPE)

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

[3]Laboratório de Inteligência Aplicada – Universidade do Vale do Itajaí (UNIVALI)

{anderson.pinheiro27,rafaelflmello,amaverick70,sebast.rogers,
guilherme.passer,periclesmiranda}@gmail.com

Abstract. *The educational forum is one of the tools present in Virtual Learning Environments (VLEs) that allows greater interaction between students and teachers. However, with the large number of students in VLEs it becamoes a challenge for teachers to have an adequate follow-up of the activities and information posted in the forums. With this, plagiarism has become a common practice in these environments. Therefore, this article presents a new measure to detect plagiarism in educational forums. This measure was evaluated in a database and was also applied in a discussion forum in a real class.*

Resumo. *O fórum educacional é uma das ferramentas presentes nos Ambientes Virtuais de Aprendizagem (AVAs) que permite ter uma maior interação entre alunos e professores. Entretanto, com a grande quantidade de alunos nos AVAs se torna um desafio para o professor ter um acompanhamento adequado das atividades e informações postadas nos fóruns. Com isso, o plágio tem se tornado uma prática comum nesses ambientes. Diante disso, este artigo apresenta uma nova medida para detecção do plágio em fóruns educacionais. Esta medida foi avaliada em uma base de dados e também foi aplicada em um fórum de discussão em uma turma real.*

1. Introdução

Com o crescente uso da tecnologia como ferramenta de apoio educacional, o uso de Ambientes Virtuais de Aprendizagem (AVAs) tem aumentado nos últimos anos [Nunes *et al.* 2012]. Estes ambientes disponibilizam várias ferramentas para melhorar a interação entre professores e alunos, alguns exemplos são: fórum, blog, wiki, entre outras. Estas ferramentas possuem um grande potencial para gerar conteúdo, o que pode ser usado para auxiliar nos processos de ensino e aprendizagem. Porém, devido a grande quantidade de interações entre os alunos e o professor, torna-se difícil para o professor avaliar e acompanhar todo o material que é disponibilizado pelos alunos [Akyuz e Kurt 2010].

O fórum é uma ferramenta que se destaca em relação à geração de conteúdo. Segundo o estudo promovido por 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%). Esta ferramenta tem uma característica importante, é nela que os alunos postam dúvidas,

DOI: 10.5753/cbie.sbie.2017.1177

comentários sobre a disciplina, outras fontes de assunto, possíveis respostas para questões levantadas pelo professor, entre outros.

Dentre as possíveis funcionalidades dos fóruns se destaca a questão da avaliação. Muitas disciplinas a distância utilizam a interação no fórum como forma de avaliação dos alunos. Contudo, com a grande quantidade de informação postada na ferramenta, muitas vezes se torna inviável para o professor, manualmente, detectar plágio nas respostas. Segundo Liu *et al.* (2007), o plágio pode ser considerado:

1. Externo, quando um aluno copia um texto da internet ou utiliza um texto de outra pessoa sem citar o autor;
2. Interno, quando os alunos copiam um texto que foi postado por outro aluno.

Existem trabalhos na literatura para identificação de plágio em: atividades educacionais [França e Soares 2012, Pertile *et al.* 2010], artigos científicos [Masic 2012] e trabalhos de conclusão de curso [Barbastefano e Souza 2007]. No entanto, quando o contexto é fóruns educacionais, a identificação de plágio se torna ainda mais difícil, devido principalmente ao tamanho do texto e por não exigir uma linguagem formal. Segundo Achananuparp, Hu e Xiajiong (2008), a variabilidade de expressão de linguagem natural faz com que seja difícil determinar sentenças semanticamente equivalentes.

Diante disso, este trabalho tem por objetivo propor uma nova medida de similaridade entre sentenças em português para detecção de plágio interno em fóruns educacionais. Esta medida extrai quatro características das sentenças e em seguida utiliza regressão linear para determinar um valor entre 0 e 1, onde 0 significa que as sentenças não apresentam similaridade e 1 significa que as sentenças são exatamente iguais. A medida foi avaliada na base disponibilizada pelo PROPOR[1] que possui 10.000 pares de sentenças com suas respectivas similaridades. A medida proposta alcançou melhores resultados do que todos os trabalhos relacionados. Além disto, ela foi aplicada em um fórum educacional utilizado em uma disciplina durante o semestre de 2017.1.

2. Trabalhos Relacionados

A base fundamental para a criação de sistemas automáticos de detecção de plágio é a criação de uma medida de similaridade que possa mensurar a relação existente entre dois textos. Calcular a similaridade entre duas sentenças é um problema desafiador, pois, as frases podem representar significados diferentes, mesmo que utilizem as mesmas palavras, ou apenas uma pequena mudança na ordem das palavras no texto pode influenciar o significado da sentença.

A similaridade textual é uma área bastante estudada e possui diversos eventos de PLN com trilhas[2] ou competições[3] específicas sobre esse tema. No entanto, a grande maioria desses eventos focam em textos em inglês. De forma inédita em 2016 a similaridade textual foi proposta para a língua portuguesa, como forma de competição, no *Workshop* de Avaliação de Similaridade Semântica e Inferência Textual (ASSIN[4]). Os trabalhos listados a seguir foram alguns dos que foram apresentados no *workshop* ASSIN 2016.

[1] http://propor2016.di.fc.ul.pt/
[2] http://pan.webis.de
[3] https://en.wikipedia.org/wiki/SemEval
[4] http://propor2016.di.fc.ul.pt/?page_id=381

Freire, Pinheiro e Feitosa (2016) propõem um framework chamado *FlexSTS*, o qual define diversos componentes a serem selecionados para o desenvolvimento de sistemas de STS (Similaridade Textual Semântica), agregando modelos e medidas de similaridade, *toolkits* e algoritmos do estado da arte. Um dos problemas deste trabalho foi a utilização da *WordNet* (um repositório de palavras onde substantivos, verbos, advérbios e adjetivos são organizados por uma variedade de relações semânticas) em inglês, onde eles traduziam as palavras do português para o inglês para poder utilizá-las. Este trabalho foi o quarto colocado na competição ASSIN.

Alves, Oliveira e Rodrigues (2016) apresentam duas abordagens distintas à tarefa de STS para a língua portuguesa na competição ASSIN: uma primeira abordagem, apelidada de Reciclagem, baseada exclusivamente em heurísticas sob redes semânticas; e uma segunda abordagem, apelidada de ASAPP, baseada em aprendizagem automática supervisionada. A abordagem Reciclagem não apresentou bons resultados e foi a quinta colocada na competição. Já a abordagem ASAPP foi a que obteve o segundo lugar na competição.

Hartmann (2016) obteve o primeiro lugar na competição apresentando uma abordagem com *feature* clássica da classe *bag-of-words*, a TF-IDF (*Term Frequency-Inverse Document Frequency*); e uma *feature* emergente, capturada por meio de *word embeddings*. O autor utiliza TF-IDF para relacionar sentenças que compartilham as mesmas palavras e *word embeddings* para capturar a sintaxe e semântica de uma palavra.

Além dos trabalhos para o Português, recentemente foi apresentada uma representação de sentença em três camadas para calcular a similaridade entre duas sentenças escritas em inglês [Ferreira *et al.* 2016]. Os autores propuseram uma matriz de similaridades para medir a similaridade entre as sentenças.

Na literatura também foram encontrados trabalhos com objetivos de criar métodos para detecção de plágio externo, ou seja, quando o aluno copia de fontes externas ao AVA. Pertile *et al.* (2010) apresentam uma modelagem de um agente detector de indícios de plágio, onde os autores apresentam um motor de busca que realiza buscas na internet a fim de encontrar documentos similares ao documento postado pelo aluno no AVA. Pertile e Medina (2011) apresentam os resultados da aplicação do método detector de indícios de plágio integrado ao Moodle. Arenhardt *et al.* (2012) apresentam um aprimoramento do método detector de indícios de plágio, onde, como diferenciais, os autores utilizam a técnica de *stemming* e é feita uma análise do texto retornado pelo motor de busca em relação aos arquivos que já estão no repositório do Moodle.

Na literatura não foi encontrado nenhum trabalho, para a língua portuguesa, cujo objetivo seja o plágio em fóruns educacionais. Desta forma, o diferencial do presente trabalho diante dos supracitados é propor uma nova medida para calcular a similaridade existente entre sentenças escritas em português e com essa medida detectar o plágio interno em fóruns educacionais. Esta medida extrai quatro características das sentenças. A primeira característica é a TF-IDF, também utilizada por Hartmann (2016). A segunda e a terceira característica são extraídas utilizando a matriz de similaridades proposta por Ferreira *et al.* (2016). A última característica é o tamanho das sentenças. As quatros características são dadas como entrada para um algoritmo de regressão linear e ao final é obtido um valor que representa a similaridade entre as sentenças. Este medida foi avaliada na base da competição ASSIN e também foi aplicada em um fórum educacional.

3. Medida de Similaridade Proposta

A medida proposta extrai 4 características dos pares de sentenças utilizando: (i) TF-IDF e similaridade de cosseno; (ii) Word2Vec e uma matriz com similaridades entre as palavras para calcular a similaridade entre as sentenças; (iii) uma matriz de similaridade binária; (iv) o tamanho das sentenças. Estas características são apresentadas a seguir.

3.1. TF-IDF

O esquema TF-IDF (*Term Frequency-Inverse Document Frequency*) é uma abordagem clássica da área de PLN. Segundo Baeza-Yates e Ribeiro-Neto (2013) TF-IDF é uma medida estatística destinada a medir o grau de importância de uma palavra para um conjunto de documentos (neste trabalho sentenças). TF-IDF combina a frequência dos termos (TF) e a relevância do termo para uma coleção (IDF). Desta forma, o esquema TF-IDF para similaridade entre sentenças é calculado pelas equações (1), (2) e (3).

$$TF = \left(\frac{número\ de\ vezes\ em\ que\ um\ termo\ aparece\ em\ dada\ sentença}{número\ total\ de\ termos\ presentes\ na\ sentença}\right) \quad (1)$$

$$IDF = 1 + log_e\left(\frac{número\ total\ de\ sentenças}{quantidade\ de\ sentenças\ que\ apresentam\ determinado\ termo}\right) \quad (2)$$

$$TF\text{-}IDF = TF * IDF \quad (3)$$

Ao final é obtida uma matriz com tamanho sentenças x palavras e o valor TF-IDF de cada palavra para cada sentença. A similaridade entre as sentenças é calculada pela distância do cosseno entre os vetores TF-IDF dos pares de sentenças. Antes de obter a matriz TF-IDF foram realizados dois pré-processamentos definidos por Hartmann (2016): para cada palavra das sentenças expandir os sinônimos para palavras que possuem até 2 sinônimos no TEP (Thesaurus para o português do Brasil) [Maziero e Pardo, 2008]; e utilizar os *stems* das palavras para reduzir a esparsidade dos dados. Foram realizados vários testes (variando a quantidade de sinônimos e modificando os pré-processamentos) e o melhor resultado foi obtido utilizando estes pré-processamentos. O valor de similaridade obtido é utilizado como característica.

3.2. Matriz de similaridades

A segunda e a terceira característica foram obtidas utilizando o método proposto por Ferreira *et al.* (2016) que utiliza uma matriz de similaridades entre palavras para obter a similaridade entre as sentenças. A similaridade entre as palavras são obtidas de duas formas e serão detalhadas nas seções seguintes (3.2.1 e 3.2.2). A seguir são mostrados os passos deste método.

O primeiro passo é calcular a similaridade entre as palavras de duas sentenças. Seja $A = \{a_1, a_2, \dots, a_n\}$ e $B = \{b_1, b_2, \dots, b_m\}$ duas sentenças, onde a_i é uma palavra da sentença A, b_j é uma palavra da sentença B, n é o número de palavras da sentença A e m é o número de palavras da sentença B. Então é calculado o valor de similaridade entre cada palavra da sentença A com cada palavra da sentença B. Por exemplo, sejam duas sentenças A e B, cada uma com seis palavras cada. São calculadas as similaridades de todas as palavras da sentença A com todas as palavras da sentença B. A Figura 1 mostra as similaridades obtidas. A maior similaridade obtida foi entre as palavras a_4 e b_6 com valor 1.0, ou seja, as duas palavras são iguais.

	a_1	a_2	a_3	a_4	a_5	a_6
b_1	0.3	0.2	0.56	0.88	0.25	0.13
b_2	0.12	0.5	0.31	0.22	0.87	0.65
b_3	0.56	0.23	0.5	0.28	0.6	0.63
b_4	0.7	0.62	0.6	0.38	0.12	0.1
b_5	0.84	0.21	0.54	0.78	0.29	0.56
b_6	0.4	0.35	0.47	**1.0**	0.23	0.33

Figura 1. Exemplo de similaridades entre palavras.

O segundo passo é remover as palavras que tiveram a maior similaridade no passo anterior. No exemplo acima seriam removidas as palavras a_4 e b_6, pois como a_4 é a mais similar com b_6 resta saber quais as palavras mais semelhantes entre as outras palavras das frases, como mostrado na Figura 2.

	a_1	a_2	a_3	a_4	a_5	a_6
b_1	0.3	0.2	0.56	0.88	0.25	0.13
b_2	0.12	0.5	0.31	0.22	0.87	0.65
b_3	0.56	0.23	0.5	0.28	0.6	0.63
b_4	0.7	0.62	0.6	0.38	0.12	0.1
b_5	0.84	0.21	0.54	0.78	0.29	0.56
~~b_6~~	~~0.4~~	~~0.35~~	~~0.47~~	~~**1.0**~~	~~0.23~~	~~0.33~~

Figura 2. Segundo passo do cálculo de similaridade entre as sentenças A e B.

Os passos 1 e 2 são repetidos até que não tenham mais palavras para calcular a similaridade. O terceiro passo é calcular a média entre os maiores valores de similaridades obtidos entre as sentenças, como mostrado na equação (4).

$$Similaridade(A,B) = \frac{\sum_{i=1}^{n} MaxSim(A,B)}{n} \quad (4)$$

Onde $MaxSim(A,B)$ é um vetor com as maiores similaridades obtidas pelos passos 1 e 2 entre as sentenças A e B, e n é o tamanho desse vetor. O valor de similaridade entre a sentença A e B será a média das maiores similaridades obtidas entre cada palavra da sentença A e cada palavra da sentença B. Neste trabalho, os valores de similaridades da matriz são obtidos de duas formas: uma utilizando o modelo Word2vec e outra definindo 1 para palavras iguais e 0 para palavras diferentes. As duas formas de obter a similaridade entre as palavras são detalhadas a seguir.

3.2.1. Matriz de similaridades com Word2vec

O modelo Word2vec surge da ideia de prever os vizinhos de uma palavra usando uma rede neural. Através do treinamento, o Word2vec simplifica o processamento do contexto para o processamento vetorial no espaço vetorial K-dimensional. Portanto, podemos obter as representações vetoriais das palavras e a similaridade entre elas pode ser calculada. O vetor de palavras pode ser considerado como um mapeamento de espaço de contexto para espaço de vetor, e pode representar fielmente as palavras [Mikolov *et al.* 2014].

O modelo Word2vec foi construído usando a implementação original[5] sobre a base da wikipedia[6] e textos de notícias obtidos pelo portal G1[7] no período de 15 de

[5] http://code.google.com/p/word2vec
[6] https://dumps.wikimedia.org/ptwiki/20160920/

Setembro a 05 de Dezembro de 2016. Etapas de pré-processamento padrão foram executados (transformar todas as letras para minúsculas e remoção de caracteres especiais). O modelo foi obtido com os seguintes parâmetros básicos: a dimensão foi definida para 250, a janela foi definida para 10, a frequência mínima de palavras foi definida para 5 e o número de iterações foram 10. Para este método foram removidas as *stopwords* e realizado *lemmatization* nas sentenças antes do cálculo da similaridade. A similaridade entre duas sentenças é obtida utilizando a matriz apresentada na Subseção 3.2 e as similaridades entre as palavras são obtidas com o Word2vec. O valor final da similaridade obtida entre os pares de sentenças é usado como característica.

3.2.2. Matriz de similaridades binária

O método anterior obtém valores de similaridades altos. Ou seja, mesmo que as sentenças apresentem um valor de similaridade baixo, a matriz de similaridades obtém um valor de similaridade alto. Isso acontece porque o método da matriz pega as maiores similaridades obtidas entre as sentenças e tira a média. Dessa forma propomos uma matriz de similaridade binária para atingir os pares de sentenças com valores de similaridades baixos. Este método também utiliza a matriz de similaridades usada no Word2vec, a diferença é que os valores de similaridade entre as palavras são obtidos pela equação (5).

$$Similaridade(a,b) = \begin{cases} 1, & se\ as\ palavras\ são\ iguais \\ 0, & se\ as\ palavras\ são\ diferentes \end{cases} \quad (5)$$

Ou seja, o valor de similaridade entre duas palavras será 1 se elas forem iguais ou 0 se elas forem diferentes. Ao final é obtida a média das similaridades entre as palavras para obter a similaridade entre as sentenças. Este valor de similaridade é usado como característica. Para este método foram removidas as *stopwords* e utilizado os *stems* das palavras.

3.3. Tamanho das sentenças

A última característica extraída, também utilizada por Zhao, Zhu e Lan (2014), foi o tamanho das sentenças. Para obter um valor que represente o tamanho das sentenças, é dividido o número de palavras da menor sentença pelo número de palavras da maior sentença. Para este método foram removidas as *stopwords*.

3.4. Regressão Linear

Para obter um valor de similaridade final entre as sentenças, a partir das características extraídas, foi utilizado o algoritmo de regressão linear. A regressão linear consiste na realização de uma análise estatística com o objetivo de verificar a existência de uma relação funcional entre uma variável dependente com uma ou mais variáveis independentes [Coelho e Cunha 2007]. As características extraídas são dadas como entrada em conjunto com uma classe (similaridade) a qual pertencem e a regressão linear obtém uma relação entre estas características que se aproxime da classe. Para este trabalho foi utilizado o algoritmo de regressão linear presente no Weka[8].

4. Experimentos

Nesta seção são descritas a base de dados e as medidas de avaliação utilizadas para avaliar a medida proposta.

[7] http://g1.com.br/
[8] https://pt.wikipedia.org/wiki/Weka

4.1. Base de Dados

Neste trabalho foi utilizada a base de dados da competição ASSIN[9]. Esta base possui 10.000 (dez mil) pares de sentenças, metade em português brasileiro e a outra metade em português europeu. Das 5.000 (cinco mil) sentenças em português brasileiro, 3.000 (três mil) são para treino e 2.000 (dois mil) são para teste. A base é dividida da mesma forma para o português europeu. Cada par de sentenças possui um valor de similaridade semântica anotado.

4.2. Medidas de Avaliação

Para a análise dos resultados foi utilizado o coeficiente de correlação de Pearson e o Erro Quadrático Médio (EQM). O coeficiente de correlação de Pearson (ρ) varia de -1 a 1. O sinal indica direção positiva ou negativa do relacionamento e o valor sugere a força da relação entre as variáveis. Quanto mais próximo de 1 for o coeficiente, maior é o grau de dependência estatística entre as variáveis [Dalson e José 2009]. O coeficiente de correlação é calculado pela equação (6). Onde x é o valor de similaridade obtido e y é o valor de similaridade desejado.

$$\rho = \frac{covariância(x,y)}{\sqrt{variância(x) * variância(y)}} \quad (6)$$

O EQM é definido como sendo a soma do quadrado das diferenças entre o conjunto de valores estimados e o conjunto de valores reais dos dados, ponderados pelo número de termos. O EQM é calculado pela equação (7).

$$EQM(Y,Y') = \sum_{i}^{n} \frac{(y_i - y'_i)^2}{n} \quad (7)$$

Onde Y é o conjunto de valores estimados, Y' é o conjunto de valores reais e n é o número de termos do conjunto. Quanto mais próximo de 0 for o erro, mais eficaz será o estimador.

5. Resultados

Os experimentos foram conduzidos na base de dados ASSIN para o Português Brasileiro (PTBR), para o Português Europeu (PTPT) e para os dois juntos (TOTAL). A Tabela 1 compara o resultado obtido com a medida proposta com os resultados obtidos pelas equipes da competição ASSIN 2016.

Tabela 1. Comparação da medida proposta com as equipes da competição ASSIN 2016.

	PTBR		PTPT		TOTAL	
Equipe/Método	**Pearson**	**EQM**	**Pearson**	**EQM**	**Pearson**	**EQM**
Medida proposta	**0,71**	**0,37**	0,70	**0,57**	**0,70**	**0,47**
Solo Queue	0,70	0,38	0,70	0,66	0,68	0,52
Reciclagem	0,59	1,31	0,54	1,10	0,54	1,23
Blue Man Group	0,65	0,44	0,64	0,72	0,63	0,59
ASAPP	0,65	0,44	0,68	0,70	0,65	0,57
LEC-UNIFOR	0,62	0,47	0,64	0,72	0,62	0,59
L2F/INESC-ID			**0,73**	0,61		

[9] http://nilc.icmc.usp.br/assin/

Como podemos perceber na Tabela 1 nossa medida obteve os melhores resultados para o PTBR e no resultado total (PTBR + PTPT), para o PTPT nosso método obteve o coeficiente de Pearson abaixo da equipe L2F/INESC-ID, mas obteve o melhor EQM. É importante destacar que a equipe L2F/INESC-ID obteve o primeiro lugar na competição para o PTPT, entretanto essa equipe não apresentou método para o PTBR. Dessa forma, a medida proposta apresentou o melhor resultado no total com uma abordagem híbrida para o português brasileiro e para o português europeu.

6. Aplicação da Medida Proposta para Detecção de Plágio Interno em Fóruns Educacionais

Com o objetivo de realizar uma avaliação da utilização da medida proposta em uma aplicação real, foi desenvolvida uma ferramenta de fórum que possui dois diferenciais, como mostrado nas Figuras 3 e 4: (i) uma barra mostrando o nível de originalidade, inversamente proporcional ao nível de similaridade entre as postagens, do fórum para os alunos; (ii) análise de estatísticas de similaridade entre postagens para o professor, onde para cada postagem aparece a postagem mais similar e a similaridade entre as postagens. É importante enfatizar que a página com as estatísticas de similaridade entre postagens é apresentada apenas para o professor (administrador do fórum criado).

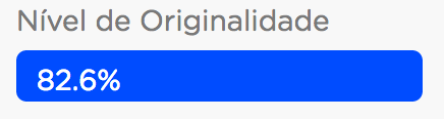

Figura 3. Barra que mostra a originalidade das postagens no fórum.

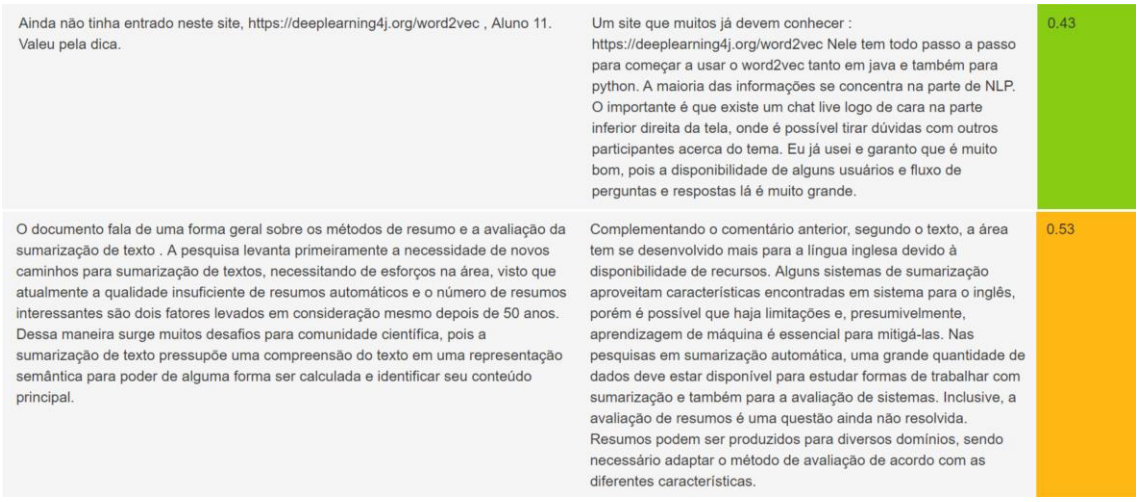

Figura 4. Página com as estatísticas de similaridade entre as postagens do fórum.

Para avaliar a eficiência da abordagem, realizou-se um experimento em uma turma de tópicos avançados em inteligência artificial que possuía 12 alunos. Foi proposta a utilização de fórum para discussão de artigos científicos relacionados ao tema da disciplina, os alunos teriam uma semana para realizar as postagens. Em um primeiro momento, foi utilizado um fórum tradicional, utilizando o Moodle, onde houve apenas 10 postagens nesse fórum.

Na atividade seguinte foi passado outro artigo para a turma, mas foi utilizada a ferramenta proposta, com estatísticas para professores e alunos. As similaridades entre as postagens foram classificadas em 4 grupos: azul (0-0,3), verde (0,31-0,5), amarelo

(0,51-0,7) e vermelho (acima de 0,7). As cores azul e verde significam que as postagens não apresentam muita semelhança. As cores amarela e vermelha significam que as postagens estão em um nível de similaridade maior, o que pode significar o plágio, principalmente a vermelha. Dessa forma, caso a postagem fosse classificada no grupo vermelho uma mensagem privada era enviada ao aluno, antes da postagem ser publicada no fórum, sugerindo que ele refizesse a postagem citando a postagem anterior (com nível de similaridade alto). Nesse fórum o número de postagem subiu para 30.

Além do acréscimo do número de postagens, na primeira semana a média de similaridade entre as postagens foi de 0,28 e com um desvio padrão de 0,13; enquanto que na segunda semana (ferramenta de fórum com a medida proposta) a média foi 0,27 e o desvio padrão 0,09. Em outras palavras, o nível de similaridade, consequentemente de plágio, foi menor com a utilização da ferramenta proposta, mesmo tendo três vezes mais postagens no fórum, e em geral as postagens tiveram um nível de similaridade mais próxima (menor desvio padrão).

Por fim, também foi relevante para o professor o fato de poder contar com as estatísticas relacionadas às postagens similares. Esse fator foi levado em consideração para a atribuição final da nota dos alunos nos fóruns.

7. Considerações Finais

Este trabalho teve por objetivo apresentar uma nova abordagem para detecção de plágio em fóruns educacionais. Para isso, uma medida de similaridade textual foi proposta. Esta medida extrai quatro características das sentenças, em seguida utiliza regressão linear para obter um valor de similaridade final entre 0 e 1.

A medida foi avaliada em uma base de dados anotada de uma competição de PLN e apresentou resultados melhores que o primeiro colocado da competição. Além disso, foi criada uma ferramenta de fórum que utiliza a medida proposta para calcular a similaridade entre as postagens. A ferramenta motiva os alunos a postarem no fórum com postagens mais originais, apresentando uma barra de originalidade, e também apresenta para o professor uma página com as estatísticas de similaridade entre as postagens do fórum.

Em trabalhos futuros pretende-se aplicar ao fórum um dos métodos de detecção de plágio externo encontrados na literatura em conjunto com a abordagem proposta neste trabalho.

8. Referências

Arenhardt, C. P. B., Medina, R. D., de Lurdes Pertile, S., Gomes, R. B., & Trindade, V. L. (2012). “Miss Marple–Proposta de Desenvolvimento de Ferramenta de Detecção de Indícios de Plágio com base no Método DIP–Detector de Indícios de Plágio” Em Anais do XXIII Simpósio Brasileiro de Informática na Educação (SBIE), v. 23, n. 1.

Akyuz, H.I. e Kurt, M. (2010) “Effect of teacher’s coaching in online discussion forums on students’ perceived self-efficacy for the educational software development” In Procedia - Social and Behavioral Sciences, v. 9, p. 633 – 637.

Alves, A. O., Oliveira, H. G. e Rodrigues, R. (2016) “ASAPP: Alinhamento Semântico Automático de Palavras aplicado ao Português” Avaliação de Similaridade Semântica e Inferência Textual.

Achananuparp, P., Hu, X. and Xiajiong, S. (2008) “The Evaluation of Sentence Similarity Measures” Lecture Notes in Computer Science, 5182. p. 305-316.

Baeza-Yates, R., e Ribeiro-Neto, B. (2013) “Recuperacao de Informacao-: Conceitos e Tecnologia das Maquinas de Busca”, Bookman Editora.

Barros, M.G. e Carvalho, A.B.G., (2011) “As concepções de interatividade nos ambientes virtuais de aprendizagem”, Tecnologias digitais na educação, p. 209-232.

Barbastefano, R. G. e Souza, C. G. (2007) “Plágio em trabalhos acadêmicos: uma pesquisa com alunos de graduação” Em XXVII Encontro Nacional de Engenharia de Produção (ENEGEP), Outubro de 2007.

Coelho, A.C., e Cunha, J.V.A, (2007) "Regressão linear múltipla." Análise multivariada: para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas (2007): 131-231.

Dalson, B., José A. (2009) "Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r)", Revista Política Hoje, Vol. 18, n. 1, pp. 115-148.

Ferreira, R., Lins, R.D., Simske, S.J., Freitas, F. e Riss, M., (2016) “Assessing sentence similarity through lexical, syntactic and semantic analysis,” Em: Computer, Speech & Language, vol. 39, Setembro 2016, p. 1-28.

França, A.B. e Soares, J.M., (2012) “Sistema de apoio a atividades de laboratório de programação com suporte ao balanceamento de carga e controle de plágio”, Em anais do XXIII Simpósio Brasileiro de Informática na Educação (SBIE), Rio de Janeiro.

Freire, J., Pinheiro, V. e Feitosa, D. (2016) “LEC_UNIFOR no ASSIN: FlexSTS - Um Framework para Similaridade Semântica Textual” Avaliação de Similaridade Semântica e Inferência Textual.

Hartmann, N. S. (2016) “Solo Queue at ASSIN: Combinando Abordagens Tradicionais e Emergentes” Avaliação de Similaridade Semântica e Inferência Textual.

Liu, Y. T., Zhang, H. R., Chen, T. W., e Teng, W. G. (2007). “Extending Web search for online plagiarism detection”. In Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on (pp. 164-169).

Maziero, E., e Pardo, T. (2008) “Interface de Acesso ao TeP 2.0 - Thesaurus para o português do Brasil,” Relatório técnico, University of São Paulo.

Mikolov, T., Chen, K., Corrado, G., Dean, J., Sutskever, L., & Zweig, G. (2014). word2vec.

Nunes, C. S., Torres, M. K. L., de Oliveira, P. C., e Nakayama, M. K. (2012). “O ambiente virtual de aprendizagem Moodle: recursos para os processos de Aprendizagem Organizacional”, Em Anais do XXIII Simpósio Brasileiro de Informática na Educação (SBIE), v. 23, n. 1.

Pertile, S.L., Piovesan, S.D., Lobo, J.S e Medina, R.D. (2010) “Agente Integrado a Plataforma MLE-Moodle para Detecção Automática de Indícios de Plágio”, Em anais do XXI Simpósio Brasileiro de Informática na Educação (SBIE), 2010.

Pertile, S.L. e Medina, R.D. (2011) “Desenvolvimento e Aplicação de um Método para Detecção de Indícios de Plágio”, Em anais do XXII Simpósio Brasileiro de Informática na Educação (SBIE), 2011, Aracajú, 1673-1682.

Zhao, J., Zhu, T.T. e Lan, M., (2014) “ECNU: One Stone Two Birds: Ensemble of Heterogenous Measures for Semantic Relatedness and Textual Entailment,” In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), August 2014, pp. 271-277.