# USO DE WORD EMBEDDINGS NA CLASSIFICAÇÃO DE REDAÇÕES POR TEMA

Rafael de Souza Freire (PIBIC/CNPq/FA/Uem),
Valéria Delisandra Feltrim (Orientadora), e-mail: ra101971@uem.br

Universidade Estadual de Maringá/Centro de Tecnologia/Maringá, PR.

**Ciência da Computação/Metodologia e Técnicas da Computação**

**Palavras-chave:** redações do ENEM**,** *embeddings*, processamento de linguagem natural

**Resumo:**

A avaliação automática de redações é um campo de estudo multidisciplinar que envolve pesquisas em diversas áreas do conhecimento. Do ponto de vista computacional, se busca construir recursos e ferramentas que automatizem, mesmo que parcialmente, a avaliação de aspectos relacionados à qualidade do texto. No caso de redações escritas como parte de exames de proficiência ou vestibulares, um desses aspectos costuma ser a pertinência do texto escrito a um tema pré-definido. Dado esse contexto, este projeto dá continuidade a uma pesquisa iniciada em 2017 que busca explorar métodos para a modelagem de tópicos aplicados à classificação de redações em temas. Atualmente, a pesquisa emprega análise de semântica latente em um sistema classificador capaz de determinar a pertinência de uma redação a um ou mais temas dados. Neste projeto o objetivo foi investigar se o uso de modelos distribuídos de língua, chamados de *embeddings*, melhoraria o desempenho da classificação. As avaliações dos modelos implementados foram realizadas utilizando um córpus de redações do ENEM. Os resultados obtidos com o modelo Word2Vec foram superiores aos obtidos com LSA, mostrando que os *embeddings* são modelos adequados para a tarefa de classificação proposta.

## Introdução

Abordagens voltadas à avaliação automática de redações se fazem cada vez mais necessárias para acompanhar o grande volume de redações circulantes em escolas e processos seletivos por todo o país. A avaliação desses textos, em especial os produzidos em provas de vestibulares, requer profissionais treinados e tempo considerável para avaliá-los, constituindo uma tarefa de alto custo. Essa demanda tem motivado a criação de recursos e ferramentas que auxiliem a avaliação de redações com um menor esforço. Este projeto se posiciona nesse cenário, focando em um aspecto específico das redações: a sua relação semântica com o tema solicitado.

A classificação de redações por temas pode ser entendida como um problema de modelagem de tópicos. Dentre os métodos tradicionalmente empregados para a

modelagem de tópicos podemos citar a análise de semântica latente (LSA) (FOLTZ et al., 2013). Recentemente, modelos de *embeddings* também têm sido aplicados a tarefas relacionadas à modelagem de tópicos.

*Embeddings* são vetores contendo números reais representando palavras (nesse caso, chamados de *word embeddings*) ou documentos em um espaço n-dimensional. Os *embeddings* normalmente são aprendidos a partir de grandes córpus e são capazes de capturar conhecimento morfológico, sintático e semântico (HARTMANN et al., 2017). Por conta do seu poder de representação semântica, esses modelos podem ser empregados para medir a similaridade entre palavras e, consequentemente, para verificar a similaridade semântica entre documentos.

O objetivo deste projeto foi avaliar se o uso de *embeddings* contribuiria para o desempenho da classificação das redações por tema. Para isso, implementou-se um sistema classificador que utilizou os diferentes modelos de *word embeddings* disponibilizados por Hartmann et al. (2017), bem como um modelo Doc2Vec. O classificador foi avaliado e seus resultados foram comparados aos de um classificador baseado em LSA.

## Materiais e métodos

O córpus de redações utilizado neste projeto foi disponibilizado por Guilherme Passero na plataforma GitHub. Esse córpus contém 2.164 redações extraídas a partir do Banco de redações da UOL distribuídas em 112 temas (com média de 19 redações por tema). O córpus está em formato XML e disponibiliza, além do texto original da redação, o título e a descrição do tema, o texto corrigido e as notas atribuídas por revisores da UOL segundo os critérios usados pelo ENEM. Antes da sua utilização, foi feita uma filtragem das redações com notas inferiores a 6,0, uma vez que redações com notas muito baixas podem fugir ao tema e distorcer os resultados de similaridade. A versão filtrada desse córpus possui 700 redações.

Para a construção do espaço semântico por meio da LSA foram utilizados, além do córpus de redações, três córpus em português disponibilizados pelo NILC-USP. São eles: G1, Wikipedia e Google News. A abordagem baseada em LSA foi avaliada em dois experimentos. No primeiro (LSA NILC), o córpus NILC foi usado para a construção do espaço semântico e o córpus de redações foi usado na avaliação. No segundo (LSA Redações), o córpus de redações foi usado tanto na construção do espaço semântico quanto na avaliação. Nos dois casos, os espaços semânticos foram gerados com 400 dimensões.

O pré-processamento dos córpus consistiu na remoção de palavras que ocorriam somente uma vez, de *stopwords* e na tokenização dos textos.

A recuperação dos *word embeddings* foi feita a partir de modelos Word2Vec disponibilizados por Hartman et al. (2017). Após experimentos preliminares, optou-se pela utilização do modelo CBOW com 1.000 dimensões.

A representação vetorial das redações se deu pela combinação dos *embeddings* recuperados para cada palavra da redação. Foram avaliados dois métodos de combinação: (1) por soma (a representação vetorial de uma redação ou tema é dada pela soma dos vetores de suas palavras) e (2) por média (uma redação ou tema é representado pela média dos vetores de suas palavras).

Também foi avaliado o modelo Doc2Vec. O modelo foi treinado a partir de um subconjunto de 10.000 textos de notícias retirados do córpus Google News. Nesse caso, utilizou-se de vetores de 100 dimensões.
Após a representação vetorial das redações (produzida por meio da LSA ou dos modelos de *embeddings*), mediu-se a similaridade do cosseno entre os vetores das redações no espaço semântico com os vetores de cada tema para analisar a proximidade de uma redação com o tema. Selecionou-se o tema de maior similaridade como resultado de classificação.
Também foi feito o agrupamento (*clustering*) das redações utilizando o algoritmo *K-Means*, com K igual ao número de temas disponíveis no córpus (112). Após o agrupamento, os grupos (*clusters*) foram analisados para verificar a predominância de um tema. Um tema foi considerado predominante quando pelo menos 40% das redações do grupo pertenciam ao tema. Grupos compostos de apenas uma redação foram descartados.
Todas as implementações foram feitas usando a linguagem de programação Python, o *framework* Gensim e a biblioteca scikit-learn.

**Resultados e Discussão**

Os resultados de classificação gerados utilizando os diferentes modelos de representação vetorial para o córpus filtrado de redações são apresentados na Tabela 1. Os modelos Word2Vec e Doc2Vec obtiveram os mesmos resultados usando tanto a combinação por soma como por média.

**Tabela 1** – Resultados de classificação obtidos com os diferentes modelos para o córpus de redações filtradas por nota

| | Total de redações comparadas | Número total de acertos | Percentual de acerto |
|---|---|---|---|
| Word2Vec | 700 | 319 | 45,57% |
| Doc2Vec | 700 | 136 | 19,46% |
| LSA Redações | 700 | 219 | 31,29% |
| LSA NILC | 700 | 124 | 17,71% |

A classificação baseada no modelo Word2Vec obteve o melhor resultado, conseguindo classificar 319 redações corretamente (45,57%). Já a classificação baseada no modelo Doc2Vec obteve um desempenho inferior, acertando somente 136 textos (19,46%). Um dos fatores que pode ter contribuído para a queda de desempenho foi a diferença do tamanho do córpus usado no treinamento dos modelos. A classificação baseada na LSA Redações obteve o segundo melhor resultado (31,29% de acerto) e a classificação baseada na LSA NILC obteve o pior desempenho, classificando corretamente apenas 124 redações (17,71%).
Com relação ao agrupamento das redações, a análise dos grupos gerados mostrou que as representações vetoriais usadas são capazes de capturar a semântica dos textos ao ponto de textos de um mesmo tema serem agrupados. No entanto, existe grande sobreposição semântica entre alguns temas, fazendo com que temas diferentes sejam considerados semanticamente similares e incluídos em um mesmo

grupo, prejudicando assim a qualidade do agrupamento por tema. Dos 112 grupos gerados pelo algoritmo *K-means* utilizando-se a representação LSA Redações, 44% dos grupos tiveram algum tema predominante (4% dos grupos foram descartados). Para a representação LSA NILC, esse número caiu para 25% (5% dos grupos foram descartados). Com relação aos modelos de *embeddings*, a combinação por média forneceu os melhores resultados de agrupamento. Quando o Word2Vec foi utilizado como representação, 58% dos grupos tiveram um tema predominante, superando os resultados obtidos com a LSA. No entanto, o número de grupos descartados foi alto (24%). O Doc2Vec teve um resultado inferior ao Word2Vec, com 52% dos grupos com um tema predominante, porém teve menos grupos descartados (13%).

## Conclusões

Neste trabalho avaliou-se o desempenho de modelos de *embeddings* na classificação das redações por tema. Para isso, implementou-se um sistema classificador baseado na similaridade de cosseno. Os resultados obtidos foram comparados aos de um classificador desenvolvido em projeto anterior baseado em LSA e mostraram que os *embeddings* foram capazes de representar melhor as redações. O modelo Word2Vec pré-treinado foi o que obteve os melhores resultados. O modelo Doc2Vec obteve resultados abaixo da LSA, possivelmente devido à falta de textos de treinamento. Também foi feito o agrupamento das por meio do algoritmo *K-means*. Os resultados indicaram que as redações de um mesmo grupo têm semântica próxima, no entanto, isso nem sempre significou que elas fossem de um mesmo tema. Independentemente do modelo de representação vetorial e do método de classificação/agrupamento utilizados, notou-se que a questão da sobreposição semântica entre os temas foi um fator dificultador.

## Agradecimentos

Ao Programa Institucional de Bolsas de Iniciação Científica – PIBIC/CNPq-Fundação Araucária-UEM pelo apoio financeiro.

## Referências

FOLTZ, P.; STREETER, L.; LOCHBAUM, K.; LANDAUER, T. Implementation and Applications of the Intelligent Essay Assessor. In: SHERMIS, M.D.; BURSTEIN, J. (Org.). **Handbook of Automated Essay Evaluation - Current Applications and New Directions**. 1st ed. New York: Taylor & Francis, 2013, p. 68–88

HARTMANN, N. S.; FONSECA, E. R.; SHULBY, C. D.; TREVISO, M. V.; RODRIGUES, J. S.; ALUÍSIO, S. M. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 2017. **Proceedings**... 2017. p.122-131