CluWords: Explorando Clusters Semânticos entre Palavras para Aprimorar Modelagem de Tópicos

Christian Gomes[1], Felipe Viegas[2], Washington Luiz[2], Leonardo Rocha[1]

[1] DCOMP/UFSJ - São João del-Rei, MG , Brasil
[2] DCC/UFMG - Belo Horizonte, MG , Brasil

{christian,lcrocha}@ufsj.edu.br,{frviegas,washington}@dcc.ufmg.br

Abstract. *In this paper, we advance the state-of-the-art in topic modeling by means of a new document representation based on pre-trained word embeddings for non-probabilistic matrix factorization. Our strategy, called CluWords, exploits the nearest words of a given pre-trained word embedding to generate meta-words capable of enhancing the document representation, in terms of both, syntactic and semantic information. In our evaluation, covering 12 datasets and 8 state-of-the-art baselines, we exceed in most cases, with gains of more than 50% against the best baselines. We also show that our method is able to improve document representation for the task of automatic text classification.*

Resumo. *Neste trabalho avançamos o estado-da-arte na modelagem de tópicos por meio de uma nova representação de documentos baseada em word embeddings pré-treinados para fatoração de matriz não-probabilística. Nossa estratégia, chamada CluWords, explora as palavras mais próximas em um determinado espaço word embedding pré-treinado para gerar meta-palavras que são capazes de melhorar a representação de documentos, tanto em termos de informações sintáticas quanto semânticas. Em nossa avaliação, considerando 12 bases de dados e 8 linhas de base, obtivemos melhoras na maioria dos casos, com ganhos de mais de 50%. Nosso método também é capaz de melhorar representação dos documentos para a tarefa de classificação automática.*

1. Introdução

As técnicas de Modelagem de Tópicos (MT) tem como objetivo descobrir padrões no uso de palavras e conectar documentos que compartilham padrões. Essas técnicas assumem que um documento é composto por tópicos ou temas, e um tópico é composto por uma coleção de palavras que o representa como um todo [Alghamdi and Alfalqi 2015]. Tratam-se, portanto, de técnicas de Aprendizado de Máquina (AM) que visam extrair tópicos "implícitos" de uma coleção de documentos e atribuir os mais relevantes para cada documento [Alghamdi and Alfalqi 2015]. Um dos principais desafios da MT está nas formas atuais com que os documentos são codificados: por exemplo, *bag-of-words* (BOW) onde cada palavra é representada pelo TF-IDF. Por meio dessa representação, a semântica dos tópicos é extraída analisando apenas as propriedades sintáticas dos dados textuais, assumindo a premissa de alta correlação entre sintaxe e semântica.

Nesse trabalho, apresentamos uma nova representação de documentos baseada em um *word embedding* pré-treinado para ser utilizado por técnicas de MT baseadas em fatoração de matriz não-probabilística (NMF). Especificamente, nossa estratégia,

chamada *CluWords (Cluster of Words)*, explora as palavras mais próximas no espaço de um *word embedding* pré-treinado para gerar "meta-palavras" que são capazes de melhorar a representação do documento, em termos de sintaxe e informação semântica. A exploração explícita da similaridade entre o *word embedding* para encontrar as palavras mais próximas fornece informações importantes sobre as relações entre as palavras. Nossa estratégia combina as evidências sintáticas tradicionais (das ocorrências de palavras em um documento) e a similaridade entre uma palavra e seus vizinhos. Essa nova representação é rica e flexível o suficiente para ser explorada por qualquer tipo de abordagem de MT. Em nossa avaliação, considerando 12 coleções de dados e 8 linhas de base, comprovamos que a estratégia proposta é mais robusta e apresenta menor variabilidade do que novas representações estado-da-arte [Shi et al. 2018]. Obtivemos melhoras em quase todos os casos, com ganhos de mais de 50% contra as melhores linhas de base. Mostramos também que nosso método é capaz de melhorar a representação de documentos para a tarefa de classificação automática de texto.

Enfatizamos que a concepção da nova representação de dados e da estratégia de ponderação, bem como todas as implementações e execuções de experimentos foram realizadas pelo aluno Christian Gomes, sob a orientação do professor Leonardo Rocha. O trabalho contou com a colaboração dos alunos de pós-graduação Felipe Viegas e Washington Luiz na concepção do ambiente experimental e nas análises de resultados.

2. Trabalhos Relacionados

A estratégia de representação de dados mais tradicional para documentos textuais é a BOW, baseada em informações simples de ocorrência de termos, codificadas pelo chamado ponderação TF-IDF (e suas variantes). Embora esta abordagem seja, de longe, a mais utilizada, ela carece de informações úteis como o contexto. Uma estratégia simples que tem sido usada pra superar são os n-grams [Cavnar et al. 1994], embora ainda seja limitado na captura de informações contextuais observadas em padrões não sequenciais. Recentemente, observamos a adoção de estratégias de representação baseadas em *word embedding*, como Word2Vec, GloVe e FastText [Mikolov et al. 2017]. Esses modelos são baseados em estatísticas de co-ocorrência de conjuntos de dados textuais, representando palavras como vetores, de modo que suas semelhanças se correlacionam com a relação semântica, explorando informações contextuais. Os modelos de previsão superam consistentemente os modelos de contagem em várias tarefas, como categorização de conceito, detecção de sinônimos e relacionamento semântico [Baroni et al. 2014].

No que se refere às técnicas de MT propriamente ditas, podemos dividí-las basicamente de acordo com o modelo utilizado, probabilísticos e não-probabilisticos. No grupo dos modelos probabilísticos, uma das primeiras técnicas propostas foi o *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003], que generaliza como $P(w|z)$, a distribuição de probabilidade sobre termos w considerando documentos pertencentes ao tópico abstrato z, é estimado. Em [Cheng et al. 2014], os autores propuseram o método *Bi-term Topic Model* (BTM) para lidar com o desafio de dispersão de dados. O BTM usa o conceito de bi-termos gerados com base em estatísticas de co-ocorrência de termos frequentes. Em [Chen and Liu 2014], os autores lidam com tópicos incoerentes através de uma técnica chamada *Lifelong Topic Model* (LTM): um método iterativo que explora dados de vários domínios de aplicação que geralmente mostram algum grau de sobreposição de informação para produzir tópicos mais coerente e confiáveis.

Recentemente, [Vorontsov and Potapenko 2015] desenvolveram a abordagem *Regularization of Topic Models* (ARTM), onde o modelo básico de pLSA [Hofmann 1999] é aumentado com regularizadores aditivos. Mais especificamente, as matrizes Φ e Θ são aprendidas maximizando uma combinação linear de $L(\Phi, \Theta)$ e r regularizadores $R_i(\Phi, \Theta), \forall i = 1, \cdots, r$, com coeficientes de regularização τ_i como mostrado na Equação 1. Embedding-based Topic Model (ETM) [Qiang et al. 2017] é outra técnica que incorpora o conhecimento externo de correlação de palavras em textos curtos para melhorar a coerência da MT. O ETM não apenas resolve o problema de informações de coocorrência de palavras muito limitadas, agregando textos curtos em longos pseudo-textos, mas também utiliza um modelo regularizado *Markov Random Field* que dá às palavras correlacionadas uma chance melhor de serem colocadas no mesmo tópico. O método FS [Guzman and Maalej 2014] é uma estratégia usada para construir tópicos com informações de sentimento. Ele extrai palavras que co-ocorrem frequentemente e infere a força do sentimento dessas palavras com base na pontuação de sentimento dos documentos em que ocorreram, aplicando o LDA nessas palavras.

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R(\Phi, \Theta),\ L(\Phi, \Theta) + R(\Phi, \Theta) \rightarrow \max(\Phi, \Theta) \tag{1}$$

Nas abordagens não-probabilísticas, onde, dado os documentos do conjunto $\mathbb{D}$ e o seu respectivo vocabulário $\mathbb{V}$, o conjunto de documentos é codificado em uma matriz esparsa $A \in \mathbb{R}^{n \times m}$, onde n é a quantidade de documentos e m é o tamanho do vocabulário, e o objetivo é decompor A em submatrizes que preservam alguma propriedade ou restrição desejada. Uma técnica de fatoração de matriz bem conhecida, que é utilizável na MT, é a *Singular Value Decomposition* (SVD). Outra estratégia amplamente utilizada é a *Non-negative Matrix Factorization* (NMF). Sob essa estratégia, a matriz A é decomposta em duas submatrizes $H \in \mathbb{R}^{n \times k}$ e $W \in \mathbb{R}^{k \times m}$, tal que $A \approx H \times W$. Nesta notação, k indica o número de fatores latentes (ou seja, tópicos), H codifica a relação entre documentos e tópicos e W codifica a relação entre tópicos e termos. A restrição imposta pelo NMF é que todas as três matrizes não possuem nenhum elemento negativo. Em [Li et al. 2017], é proposto um modelo chamado GPU-DMM, que pode promover palavras semanticamente relacionadas usando as informações fornecidas pelo *word embeddings* dentro de qualquer tópico. O GPU-DMM estende o modelo de *Dirichlet Multinomial Mixture* (DMM), incorporando o aprendizado de palavras aprendidas de *word embeddings* através do modelo generalizado Pólya urn (GPU) [Mahmoud 2008] em inferências de tópicos. Por fim, em [Shi et al. 2018], os autores propõem um modelo de fatoração da matriz resultante do método NMF auxiliado pela semântica – *Semantics-Assisted Non-negative Matrix Factorization* (SeaNMF) – para descoberta de tópicos em documentos curtos. Basicamente, o método incorpora as correlações semânticas do contexto de palavras do modelo. As correlações semânticas entre as palavras e seus contextos são aprendidas a partir da visão *skip-gram* do *corpus*.

Não encontramos trabalhos que combine as informações de modelos de *word embedding* pré-treinados e modelos não probabilísticos uma vez que a introdução da representação *word embedding* dificulta a representação de tópicos devido à falta de correspondência direta entre tópicos e unidades semânticas menores (*e.g.*, palavras).

3. Estratégia Proposta

Nossa estratégia consiste em adaptar a representação tradicional de BOW para incluir informações semânticas relacionadas as palavras presentes nos documentos. Sendo

assim, considere o vocabulário $\mathbb{V}$ das palavras presentes no conjunto de documentos $\mathbb{D}$. Seja $\mathbb{W}$ o conjunto de vetores representando cada palavra em $\mathbb{V}$ de acordo com o modelo *word embedding* pré-treinado. Cada palavra w em $\mathbb{V}$ tem um vetor correspondente em $\mathbb{W}$ e cada vetor $\vec{w}, \vec{w'} \in \mathbb{W}$ tem comprimento l, onde l é a dimensionalidade do espaço vetorial do modelo. Definimos as CluWords como uma matriz $C \in \mathbb{R}^{|\mathbb{V}|\times|\mathbb{V}|}$, onde cada índice $C_{w,w'}$ é calculado de acordo com a Equação 2.

$$C_{w,w'} = \begin{cases} \omega(w,w') & \text{se } \omega(w,w') \geq \alpha \\ 0 & \text{caso contrário} \end{cases} \quad (2) \qquad \omega(u,v) = \frac{\sum_i^l u_i \cdot v_i}{\sqrt{\sum_i^l u_i^2} \cdot \sqrt{\sum_i^l v_i^2}} \quad (3)$$

onde $\omega(w, w')$ é a distância de cosseno definida na Equação 3 e α é um limiar de similaridade que controla a inclusão do valor da similaridade entre as palavras w e w'. Cada CluWord é representada como uma linha C_w e cada coluna $w' \in \mathbb{V}$ mapeia a vizinhança de C_w, atribuindo à posição C_w, w' a similaridade de cosseno $\omega(w, w')$, caso $\omega(w, w') \leq \alpha$. Caso contrário, é atribuído o valor zero. A CluWord C_w relaciona w com suas palavras mais próximas, limitando essa relação com o valor de corte α, que filtra as palavras que não possuem uma relação semântica com w. Como o limite α é um valor de similaridade de cosseno, ele está contido no intervalo $[0, 1]$. Assim, a seleção apropriada de um valor para o parâmetro α é um aspecto importante para a construção das CluWords. Pois, o α controla a qualidade da representação proposta nos documentos da coleção de dados. Por exemplo, dado um documento d, se for atribuído altos valores em α, apenas um pequeno conjunto de CluWords estarão relacionadas com o documento d. Isto ocorre devido a baixa cobertura de vizinhança semântica entre palavras, ou seja, se $\alpha \approx 1.0$, mais próximo será a representação das CluWords com a representação BOW. No entanto, atribuindo baixos valores em α, um maior conjunto de CluWords estarão relacionadas ao documento d. Contudo, valores de $\alpha \approx 0.0$ adicionarão ruído à representação das CluWords. Após realizado a seleção da vizinhança de cada CluWord, é calculado o TF-IDF.

Como podemos observar, as CluWords são criadas com base na semelhança semântica das palavras, de modo que a métrica TF-IDF convencional não é capaz de ponderar esses recursos. Nossa motivação é combinar os dois aspectos da métrica TF-IDF convencional (TF–relevância da palavra em um documento e IDF–a importância da palavra na coleção de documentos) com a informação semântica das CluWords. A seguir, propomos uma versão modificada do TF-IDF. O TF-IDF de uma CluWord para um documento d é definido de acordo com $C_{TF-IDF} = C_{TF} \times idf(C)$. Primeiro, o TF pode ser representado como uma matriz $T \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{V}|}$, onde cada posição $T_{d,w}$ considera a frequência de uma palavra w no documento d. O TF das CluWords pode ser medido como um produto de matrizes como descrito na Equação 4. O valor de $C_{TF_{d,w}}$ corresponde à soma dos produtos das frequências das palavras $T_{d,w'}$ de cada palavra $w' \in Cw, w' \neq 0$ ocorrendo no documento d. Para calcular o IDF de uma CluWord C_w, primeiro definimos o vocabulário $\mathcal{V}_{d,w}$ composto por todas as palavras no documento d que têm o peso ω_w diferente de zero na CluWord C_w. Isso é formalmente definido na Equação 5.

$$C_{TF} = C \times T^T \quad (4) \qquad \mathcal{V}_{d,C_w} = \{w' \in d | C_{w,w'} \neq 0 \text{ in } C_w\} \quad (5)$$

Em seguida, calculamos a média dos valores dos pesos da CluWord C_w das palavras que ocorrem no vocabulário $\mathcal{V}_{d,C_w}$, de acordo com a Equação 6. Finalmente calculamos o IDF da CluWord C_w como definido na Equação 7, onde $\mathcal{D}$ é o conjunto de treinamento.

$$\mu_{C_w,d} = \frac{1}{|\mathcal{V}_{d,C_w}|} \cdot \sum_{w \in \mathcal{V}_{d,C_w}} w_w \quad (6) \qquad idf(C_w) = log\left(\frac{|\mathcal{D}|}{\sum_{1 \leq d \leq |\mathcal{D}|} \mu_{C_w,d}}\right) \quad (7)$$

4. Avaliação Experimental

4.1. Configuração Experimental

A - Coleção de Dados: O objetivo da nossa solução é executar efetivamente a MT para que tópicos mais coerentes sejam extraídos. Para avaliar a qualidade do nosso modelo, consideramos 12 coleções de dados: duas delas criadas por nós contendo comentários de aplicativos na Google Play Store (i.e. Facebook e Uber) e outras 10 coleções amplamente utilizadas em trabalhos anteriores na literatura [Viegas et al. 2015, Guzman and Maalej 2014, Li et al. 2016] (20NewsGroup, ACM, Angrybirds, Dropbox, Evernote, InfoVisVast, Pinterest, TripAdvisor, Tweets e WhatsApp). Para todas elas, realizamos a remoção de palavras irrelevantes (usando a lista SMART padrão) e removemos palavras como advérbios, usando o dicionário VADER, pois a grande maioria das palavras importantes para identificar tópicos são substantivos e verbos.

B - Avaliação, Algoritmos e Procedimentos: Consideramos três comprimentos dos tópicos (5, 10 e 20 palavras), comparando as estratégias usando duas métricas da literatura que avaliam a qualidade representativa dos tópicos [Shi et al. 2018, Nikolenko et al. 2017]:

- *TF-IDF Coherence:* captura a facilidade de interpretação de acordo com a coocorrência das palavras [Nikolenko et al. 2017] e é definida na Equação 8.

$$c_{\text{tf-idf}}(t, W_t) = \sum_{w_1, w_2 \in W_t} log \frac{\sum_{d: w_1, w_2 \in d} \text{tf-idf}(w_1, d) \times \text{tf-idf}(w_2, d)}{\sum_{d: w_1 \in d} \text{tf-idf}(w_1, d)} \tag{8}$$

  onde a métrica *tf-idf* é dada pela frequência aumentada de acordo com a Equação 9.

$$\text{tf-idf}(w, d) = \left(\frac{1}{2} + \frac{f(w, d)}{max_{w' \in d} f(w', d)}\right) \times log\left(\frac{|D|}{|\{d \in D : w \in d\}|}\right) \tag{9}$$

  e $f(w, d)$ é o número de ocorrências de uma palavra w no documento d.
- *Normalized Pairwise pointwise Mutual Information (NPMI):* mede quanto uma palavra "ganha" de informação dada à ocorrência de outra palavra, levando em consideração as dependências entre as palavras [Nikolenko 2016]. Para um determinado conjunto ordenado das palavras mais importantes $W_t = (w_1, ..., w_N)$ de um tópico, a métrica NPMI é calculada de acordo com a Equação 10.

$$\text{NPMI}_t = \sum_{i<j} \frac{log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}}{-log\, p(w_i, w_j)} \tag{10}$$

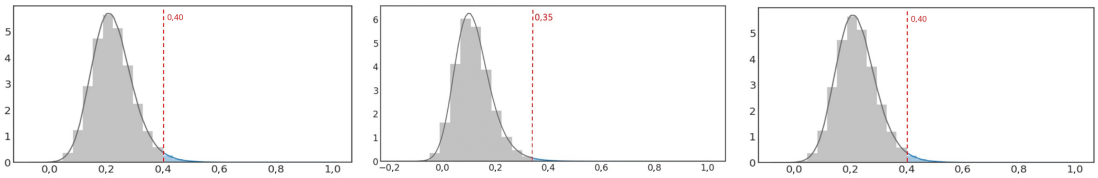
Utilizamos a técnica NMF para avaliar as CluWords, por ser a principal técnica de fatoração de matriz não probabilística. O número de tópicos para as coleções de dados de aplicativos foi definido com base nas escolhas feitas em [Guzman and Maalej 2014] (25 tópicos). Para as coleções 20News, ACM e Tweets, escolhemos o número de tópicos igual ao número de classes de cada coleção (20, 11 e 6 tópicos, respectivamente). Avaliamos a significância estatística de nossos resultados por meio do *t-test* pareado com 95% de confiança e a correção de Holm-Bonferroni para contabilizar múltiplos testes.

4.2. Resultados Experimentais

A - Escolhendo o Espaço Word Embedding: Comparamos a técnica proposta com três modelos *word embedding* [Mikolov et al. 2017] pré-treinados: (i) Word2Vec – modelo treinado com o GoogleNews; (ii) FastText – modelo treinado com WikiNews e (iii) FastText – modelo treinado em *Common Crawl*. Inicialmente, para construir a

representação de dados proposta, precisamos selecionar um limite α que seja restritivo, capaz de filtrar pares de palavras ruidosas. Para isso, precisamos encontrar a distribuição das semelhanças entre os pares de palavras no espaço *word embedding* para inferir um limiar de similaridade. A Figura 1 mostra a distribuição das similaridades de cada espaço *word embedding* pré-treinado. Podemos observar que a distribuição de similaridades nos três espaços vetoriais é bastante semelhante e que o FastText WikiNews apresenta um desvio um pouco maior que os outros modelos. Assim, para nossos experimentos, escolhemos um limite α capaz de selecionar apenas 2% dos pares de palavras similares. O limite selecionado para a FastText WikiNews é de $\alpha \geq 0,40$, enquanto para W2V GoogleNews e FastText Common Crawl, um limite de $\alpha \geq 0,35$ foi selecionado.

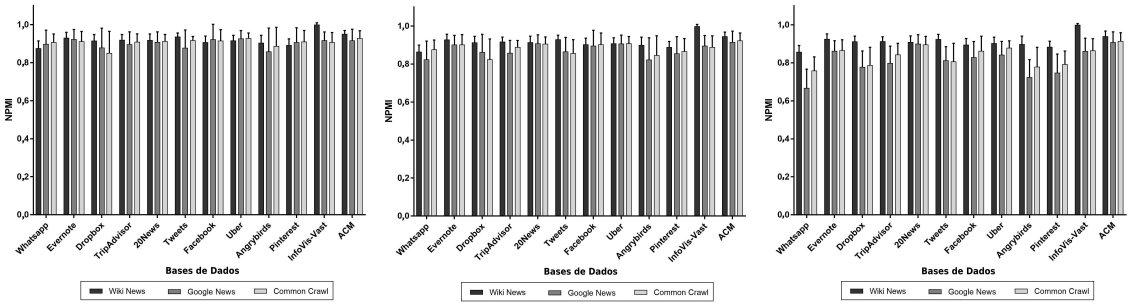
Figura 1. Histograma de similaridades de cada espaço *word embedding*.



(a) Word2Vec Google News (b) FastText WikiNews (c) FastText Common Crawl

A Figura 2 mostra os resultados das CluWords nos três espaços *word embedding* avaliados. O FastText WikiNews sempre alcança resultados superiores considerando todas as coleções de dados e tamanhos de tópicos. De fato, a maioria dos resultados (32 dos 36 resultados) apresentam um empate estatístico, o que sugere que a representação de dados proposta é capaz de realizar a MT nos três espaços *word embedding* com a mesma qualidade. Os resultados das CluWords apresentados na próxima Seção foram gerados usando o espaço *word embedding* FastText WikiNews.

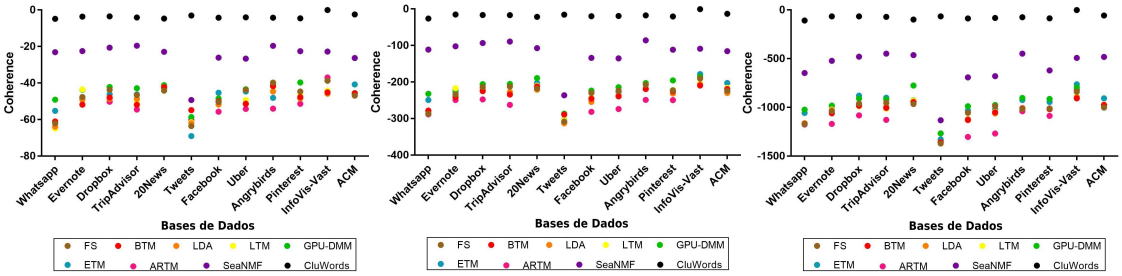
Figura 2. Avaliação das CluWords explorando diferentes *word embeddings*.



(a) Avaliação com 5 palavras. (b) Avaliação com 10 palavras. (c) Avaliação com 20 palavras.

B - Resultados de Eficácia em Comparação com as Linhas de Base: Comparamos nossa solução proposta com 8 estratégias de MT no estado da arte descritas na Seção 2 (LDA, BTM, LTM, ARTM, ETM, FS, GPU-DMM e SeaNMF). Na Figura 3, nossa estratégia obtém ganhos estatisticamente significativos em termos da qualidade dos tópicos descobertos nas 12 coleções de dados, considerando a métrica *Coherence*. A maioria das linhas de base não apresentam resultados nem perto do obtido com as CluWord, sendo que, as CluWords superam o SeaNMF (considerado, até então, o mais eficaz da literatura) em mais de 33% dos casos.

Figura 3. Comparando os resultados obtidos por cada estratégia, considerando as 5, 10 e 20 palavras principais para a métrica *TF-IDF Coherence*



(a) Resultados com 5 palavras. (b) Resultados com 10 palavras. (c) Resultados com 20 palavras.

Na Tabela 1, mostramos os resultados das CluWords e as estratégias de comparação escolhidas, considerando a métrica NPMI. Os melhores resultados, marcados com ▲, são estatisticamente superiores aos outros. Os empates estatísticos são representados por ●. Como podemos ver, nossa estratégia atinge os melhores resultados em 7 dos 36 resultados, empatando com o SeaNMF nos outros 29 casos como o método **melhor** em termos de qualidade dos tópicos descobertos. Novamente, os resultados das outras linhas de base estão muito inferiores, reforçando que o SeaNMF era a linha de base a ser vencida.

Tabela 1. Comparação dos resultados obtidos por cada estratégia, considerando as 5, 10 e 20 palavras principais para o NPMI.

Estratégia	Whatsapp			Evernote			Dropbox		
	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras
FS	0,171 ± 0,051	0,201 ± 0,048	0,230 ± 0,043	0,102 ± 0,052	0,090 ± 0,020	0,100 ± 0,018	0,109 ± 0,042	0,097 ± 0,027	0,107 ± 0,018
BTM	0,201 ± 0,057	0,236 ± 0,038	0,284 ± 0,038	0,118 ± 0,057	0,109 ± 0,029	0,120 ± 0,024	0,155 ± 0,050	0,161 ± 0,043	0,166 ± 0,040
LDA	0,172 ± 0,050	0,230 ± 0,030	0,284 ± 0,042	0,114 ± 0,067	0,114 ± 0,036	0,114 ± 0,019	0,165 ± 0,110	0,149 ± 0,056	0,150 ± 0,037
LTM	0,178 ± 0,052	0,225 ± 0,041	0,269 ± 0,040	0,193 ± 0,051	0,168 ± 0,044	0,158 ± 0,033	0,167 ± 0,072	0,160 ± 0,040	0,175 ± 0,046
GPU-DMM	0,312 ± 0,165	0,327 ± 0,141	0,330 ± 0,131	0,258 ± 0,165	0,270 ± 0,149	0,229 ± 0,076	0,284 ± 0,147	0,267 ± 0,125	0,284 ± 0,129
ETM	0,365 ± 0,171	0,378 ± 0,163	0,399 ± 0,154	0,319 ± 0,138	0,320 ± 0,133	0,331 ± 0,131	0,403 ± 0,094	0,399 ± 0,109	0,398 ± 0,119
ARTM	0,174 ± 0,046	0,248 ± 0,036	0,339 ± 0,042	0,125 ± 0,050	0,118 ± 0,019	0,139 ± 0,013	0,158 ± 0,041	0,183 ± 0,036	0,239 ± 0,030
SeaNMF	0,884 ± 0,256 ●	0,803 ± 0,166 ●	0,576 ± 0,112	0,932 ± 0,293 ●	0,901 ± 0,283 ●	0,780 ± 0,241 ●	0,968 ± 0,222 ●	0,927 ± 0,219 ●	0,784 ± 0,185 ●
CluWords	0,875 ± 0,039 ●	0,864 ± 0,036 ●	0,856 ± 0,036 ▲	0,929 ± 0,031 ●	0,928 ± 0,029 ●	0,924 ± 0,029 ●	0,914 ± 0,034 ●	0,912 ± 0,033 ●	0,912 ± 0,029 ●
Estratégia	**TripAdvisor**			**20News**			**Tweets**		
	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras
FS	0,094 ± 0,037	0,092 ± 0,028	0,104 ± 0,021	0,119 ± 0,056	0,110 ± 0,026	0,110 ± 0,022	0,071 ± 0,054	0,066 ± 0,033	0,078 ± 0,005
BTM	0,130 ± 0,052	0,144 ± 0,044	0,158 ± 0,039	0,244 ± 0,117	0,217 ± 0,089	0,192 ± 0,059	0,142 ± 0,061	0,100 ± 0,026	0,095 ± 0,019
LDA	0,114 ± 0,057	0,122 ± 0,029	0,137 ± 0,028	0,218 ± 0,121	0,196 ± 0,084	0,174 ± 0,063	0,083 ± 0,055	0,060 ± 0,028	0,079 ± 0,020
LTM	0,149 ± 0,059	0,144 ± 0,035	0,161 ± 0,037	0,224 ± 0,134	0,196 ± 0,074	0,179 ± 0,049	0,109 ± 0,060	0,084 ± 0,022	0,093 ± 0,017
GPU-DMM	0,286 ± 0,209	0,253 ± 0,144	0,244 ± 0,122	0,421 ± 0,044	0,477 ± 0,044	0,471 ± 0,031	0,090 ± 0,062	0,081 ± 0,051	0,092 ± 0,046
ETM	0,347 ± 0,151	0,349 ± 0,154	0,355 ± 0,163	0,249 ± 0,109	0,262 ± 0,092	0,243 ± 0,066	0,057 ± 0,044	0,071 ± 0,038	0,092 ± 0,041
ARTM	0,128 ± 0,042	0,168 ± 0,030	0,226 ± 0,030	0,281 ± 0,105	0,235 ± 0,076	0,216 ± 0,062	0,091 ± 0,055	0,068 ± 0,031	0,080 ± 0,025
SeaNMF	0,951 ± 0,292 ●	0,938 ± 0,293 ●	0,816 ± 0,262 ●	0,897 ± 0,247 ●	0,893 ± 0,249 ●	0,891 ± 0,254 ●	0,237 ± 0,183	0,239 ± 0,145	0,195 ± 0,056
CluWords	0,918 ± 0,030 ●	0,916 ± 0,026 ●	0,912 ± 0,025 ●	0,917 ± 0,034 ●	0,913 ± 0,034 ●	0,908 ± 0,034 ●	0,935 ± 0,021 ▲	0,928 ± 0,024 ▲	0,923 ± 0,027 ▲
Estratégia	**Facebook**			**Uber**			**Angrybirds**		
	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras
FS	0,061 ± 0,065	0,054 ± 0,033	0,050 ± 0,014	0,056 ± 0,043	0,045 ± 0,023	0,048 ± 0,016	0,053 ± 0,036	0,077 ± 0,033	0,124 ± 0,028
BTM	0,137 ± 0,063	0,110 ± 0,036	0,118 ± 0,029	0,093 ± 0,044	0,094 ± 0,036	0,094 ± 0,026	0,132 ± 0,075	0,154 ± 0,034	0,193 ± 0,040
LDA	0,115 ± 0,067	0,085 ± 0,028	0,095 ± 0,023	0,094 ± 0,053	0,083 ± 0,030	0,089 ± 0,012	0,137 ± 0,065	0,154 ± 0,038	0,190 ± 0,044
LTM	0,146 ± 0,079	0,113 ± 0,048	0,119 ± 0,027	0,097 ± 0,065	0,088 ± 0,032	0,091 ± 0,022	0,117 ± 0,061	0,154 ± 0,041	0,189 ± 0,041
GPU-DMM	0,326 ± 0,170	0,313 ± 0,164	0,282 ± 0,162	0,322 ± 0,241	0,275 ± 0,199	0,240 ± 0,142	0,260 ± 0,173	0,286 ± 0,141	0,301 ± 0,142
ETM	0,198 ± 0,090	0,186 ± 0,087	0,171 ± 0,095	0,180 ± 0,077	0,173 ± 0,074	0,165 ± 0,096	0,366 ± 0,089	0,373 ± 0,079	0,385 ± 0,085
ARTM	0,079 ± 0,044	0,091 ± 0,023	0,136 ± 0,021	0,075 ± 0,043	0,091 ± 0,020	0,135 ± 0,018	0,209 ± 0,066	0,262 ± 0,054	0,337 ± 0,051
SeaNMF	0,718 ± 0,410 ●	0,655 ± 0,396 ●	0,546 ± 0,312	0,684 ± 0,434 ●	0,630 ± 0,417 ●	0,522 ± 0,343	0,964 ± 0,238 ●	0,955 ± 0,222 ●	0,808 ± 0,194 ●
CluWords	0,917 ± 0,034 ●	0,913 ± 0,034 ●	0,908 ± 0,034 ●	0,935 ± 0,021 ▲	0,928 ± 0,024 ▲	0,923 ± 0,027 ▲	0,903 ± 0,041 ●	0,899 ± 0,043 ●	0,897 ± 0,044 ●
Estratégia	**Pinterest**			**InfoVis-Vast**			**ACM**		
	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras
FS	0,102 ± 0,077	0,096 ± 0,050	0,112 ± 0,031	0,049 ± 0,039	0,057 ± 0,026	0,056 ± 0,019	0,148 ± 0,107	0,136 ± 0,050	0,128 ± 0,044
BTM	0,148 ± 0,074	0,144 ± 0,043	0,147 ± 0,032	0,193 ± 0,079	0,170 ± 0,071	0,149 ± 0,051	0,176 ± 0,084	0,146 ± 0,055	0,136 ± 0,051
LDA	0,144 ± 0,062	0,135 ± 0,043	0,147 ± 0,039	0,154 ± 0,075	0,153 ± 0,064	0,139 ± 0,051	0,138 ± 0,062	0,122 ± 0,051	0,117 ± 0,046
LTM	0,145 ± 0,061	0,137 ± 0,051	0,143 ± 0,035	0,182 ± 0,092	0,158 ± 0,058	0,131 ± 0,042	0,173 ± 0,095	0,163 ± 0,074	0,143 ± 0,054
GPU-DMM	0,330 ± 0,192	0,322 ± 0,194	0,278 ± 0,146	0,264 ± 0,155	0,259 ± 0,104	0,207 ± 0,103	0,233 ± 0,101	0,208 ± 0,113	0,189 ± 0,095
ETM	0,358 ± 0,114	0,355 ± 0,109	0,370 ± 0,115	0,304 ± 0,163	0,304 ± 0,157	0,313 ± 0,158	0,266 ± 0,086	0,230 ± 0,052	0,201 ± 0,035
ARTM	0,167 ± 0,060	0,198 ± 0,030	0,264 ± 0,027	0,102 ± 0,095	0,084 ± 0,077	0,076 ± 0,045	0,178 ± 0,088	0,147 ± 0,058	0,149 ± 0,047
SeaNMF	0,836 ± 0,311 ●	0,754 ± 0,278 ●	0,552 ± 0,167	0,861 ± 0,321 ●	0,840 ± 0,332 ●	0,768 ± 0,288	0,843 ± 0,336 ●	0,857 ± 0,337 ●	0,860 ± 0,345 ●
CluWords	0,891 ± 0,034 ●	0,888 ± 0,031 ●	0,883 ± 0,031 ▲	0,998 ± 0,012 ●	0,997 ± 0,012 ●	0,998 ± 0,009 ▲	0,950 ± 0,019 ●	0,945 ± 0,023 ●	0,939 ± 0,028 ●

Para melhor quantificar a eficácia das CluWords, realizamos dois testes de variabilidade para comprovar que, de acordo com os desvios padrão, as CluWords apresentam

um melhor resultado do que o SeaNMF. Variações iguais entre amostras também são chamadas de *homogeneidade de variância*. Alguns testes estatísticos (*e.g.*, análise de variância) assumem que as variações são iguais entre os grupos ou amostras. O teste de Levene e o teste de Bartlett podem ser usados para verificar esta suposição. O teste de Levene é menos sensível do que o teste de Bartlett para amostras anormais. Por outro lado, se os dados vierem de fato de uma distribuição normal ou quase normal, o teste de Bartlett deve ter um desempenho melhor. Como não podemos assumir nenhuma das opções, aplicamos os dois testes. Nesses testes, se o p-value resultante for menor do que algum nível de significância as diferenças obtidas nas variâncias das amostras provavelmente não ocorreram com base na amostragem aleatória de uma população com variâncias iguais.

A Tabela 2 apresenta o teste para igualdade de variâncias com relação aos valores da métrica NPMI para as CluWords e o SeaNMF. Marcamos em ▲ os p-valores que apresentam diferenças estatisticamente significativas entre as variâncias e usamos ● para quando as duas estratégias têm a mesma variação. Em 21 dos 24 testes, as CluWods e o SeaNMF possuem variâncias diferentes. Assim, podemos concluir que nossa estratégia é capaz de gerar os melhores tópicos semanticamente coesos, em termos das métricas *TF-IDF Coherence* e NPMI de acordo com os testes de Levene e Barlett.

Tabela 2. Teste de igualdade de variâncias considerando 20 palavras.

Base de Dados	Variância		p-valor	
	CluWords	SeaNMF	Teste Levene	Teste Bartlett
20News	0,0013	0,0644	0,004▲	0.0 ▲
ACM	0,0008	0,0741	0,169●	0.018▲
Angrybirds	0,0020	0,0375	0,002▲	0.014▲
Dropbox	0,0009	0,0343	0,006▲	0.006▲
Evernote	0,0009	0,0582	0,015▲	0.000▲
Facebook	0,0012	0,0971	0,000▲	0.000▲
Infovisvast	0,0001	0,0831	0,000▲	0.000▲
Pinterest	0,0010	0,0278	0,002▲	0.001▲
TripAdvisor	0,0007	0,0687	0,004▲	0.000▲
Tweets	0,0009	0,0032	0,112●	0.138●
Uber	0,0011	0,1179	0,000▲	0.000▲
WhatsApp	0,0013	0,0125	0,001▲	0.000▲

C - Aplicação: Classificação de Documentos: Como vimos, nosso método proposto é capaz de gerar tópicos mais coesos e melhores representações de documentos, o que pode ajudar consideravelmente em tarefas como classificação automática e agrupamento. Assim, analisamos a adequação das informações do nosso modelo na tarefa de classificação, deixando a análise de outras aplicações para trabalhos futuros. Consideramos as bases de dados ACM e 20News, que já possuem uma classificação verdadeira. Comparamos três tipos de informações extraídas do modelo de tópico: (1) Tópicos latentes gerados pela estratégia CluWords; (2) Tópicos latentes gerados pelo SeaNMF e (3) A representação dos documentos pela técnica CluWord. Cada tipo de informação é combinada com a representação BOW original, que também é uma linha de base.

Todos os experimentos foram executados usando a técnica de validação cruzada com 5 conjuntos, sendo que, para a classificação foi utilizado o SVM, que é um método de alta qualidade na classificação de textos. O parâmetro de regularização foi escolhido entre onze valores de 2^{-5} a 2^{15}, usando uma validação cruzada aninhada de 5 vezes dentro

do conjunto de treinamento. Avaliamos a significância estatística de nossos resultados por meio de um *t-test* pareado com 95% de confiança e correção de Holm-Bonferroni para contabilizar vários testes. Este teste assegura que os melhores resultados, marcados com ▲, são estatisticamente superiores aos outros.

Tabela 3. Média das métricas Macro-F1 e Micro-F1 para a tarefa de classificação usando diferentes representações de documento.

Representação	ACM		20News	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
BOW	69,1 ± 0,4	57,3 ± 1,64	89,6 ± 0,5	89,5 ± 0,5
CluWords	74,0 ± 0,8	61,9 ± 1,8	91,1 ± 0,8	91,0 ± 0,9
Tópicos CluWords	76,0 ± 0,5 ▲	62,8 ± 1,5 ▲	92,4 ± 0,2 ▲	92,2 ± 0,3 ▲
Tópicos SeaNMF	71,2 ± 0,8	61,3 ± 1,4	87,0 ± 0,3	87,0 ± 0,2

A Tabela 3 apresenta a eficácia da classificação utilizando as métricas Micro-F1 e Macro-F1. Em todas situações, o uso dos tópicos latentes da técnica CluWord, e as CluWords individualmente, obtiverem melhores resultados que as outras abordagens, com significância estatística nas bases avaliadas. Esses resultados indicam que as informações fornecidas pelas CluWords podem melhorar os resultados da classificação automática de documentos, seja individualmente ou com a estratégia de MT.

5. Conclusão e Trabalhos Futuros

Neste artigo, apresentamos uma nova representação de documentos para Modelagem de Tópicos – CluWords, que é capaz de: (i) explorar relações semânticas explícitas entre palavras, sem limitações de escalabilidade e adaptabilidade; (ii) conjugar em uma única representação informações sintáticas e semânticas; e (iii) propor uma maneira de ponderar a importância das CluWords para expressar os tópicos de um documento. Nossa avaliação experimental mostrou que superamos os melhores métodos para Modelagem de Tópicos conhecidos na literatura, com grandes margens de diferença e uma variabilidade muito menor em termos de qualidade dos tópicos produzidos, se estabelecendo a nova estratégia a ser superada. Também demonstramos que os tópicos gerados têm o potencial de melhorar outras aplicações, como a classificação automática de documentos. Como trabalho futuro, precisamos compreender melhor as propriedades teóricas dos *clusters* de uma CluWord. Pretendemos também explorar novas medidas de similaridade adaptadas para particularidades de cada base de dados, tais como densidade e número de atributos.

Referências

Alghamdi, R. and Alfalqi, K. (2015). A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1).

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 238–247.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Cavnar, W. B., Trenkle, J. M., et al. (1994). N-gram-based text categorization. *Ann arbor mi*, 48113(2):161–175.

Chen, Z. and Liu, B. (2014). Topic modeling using topics from many domains, lifelong learning and big data. In *International Conference on Machine Learning*, pages 703–711.

Cheng, X., Yan, X., Lan, Y., and Guo, J. (2014). Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge & Data Engineering*, (1):1–1.

Guzman, E. and Maalej, W. (2014). How do users like this feature? a fine grained sentiment analysis of app reviews. In *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*, pages 153–162. IEEE.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., and Ma, Z. (2017). Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems (TOIS)*, 36(2):11.

Li, Q., Shah, S., Liu, X., Nourbakhsh, A., and Fang, R. (2016). Tweetsift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2429–2432. ACM.

Mahmoud, H. (2008). *Pólya urn models*. Chapman and Hall/CRC, 1 edition.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2017). Advances in pre-training distributed word representations. *CoRR*, abs/1712.09405.

Nikolenko, S. I. (2016). Topic quality metrics based on distributed word representations. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval*, pages 1029–1032. ACM.

Nikolenko, S. I., Koltcov, S., and Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, 43(1):88–102.

Qiang, J., Chen, P., Wang, T., and Wu, X. (2017). Topic modeling over short texts by incorporating word embeddings. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 363–374. Springer.

Shi, T., Kang, K., Choo, J., and Reddy, C. K. (2018). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1105–1114. International World Wide Web Conferences Steering Committee.

Viegas, F., Gonçalves, M. A., Martins, W., and Rocha, L. (2015). Parallel lazy semi-naive bayes strategies for effective and efficient document classification. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1071–1080. ACM.

Vorontsov, K. and Potapenko, A. (2015). Additive regularization of topic models. *Machine Learning*, 101(1-3):303–323.