UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Christian Reis Fagundes Gomes

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

São João del-Rei

2018

UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Christian Reis Fagundes Gomes

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

Monografia apresentada como requisito da disciplina de Projeto Orientado em Computação II do Curso de Bacharelado em Ciência da Computação da UFSJ.

Orientador: Leonardo Chaves Dutra da Rocha

Universidade Federal de São João del-Rei – UFSJ

Bacharelado em Ciência da Computação

São João del-Rei

2018

Christian Reis Fagundes Gomes

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

Monografia apresentada como requisito da disciplina de Projeto Orientado em Computação II do Curso de Bacharelado em Ciência da Computação da UFSJ.

Trabalho aprovado.
São João del-Rei, 9 de novembro de 2018:

Leonardo Chaves Dutra da Rocha
Orientador

Edimilson Batista dos Santos
Convidado 1

Diego Roberto Colombo Dias
Convidado 2

São João del-Rei
2018

Este trabalho é dedicado aos meus pais Andrea e Carlos,
à minha falecida avó Maria Helena,
à minha irmã Carol, à minha namorada Ayeska e aos
meus colegas e professores que me apoiaram
em minha trajetória.

Agradecimentos

Dedico esta, bem como todas as minhas demais conquistas, aos meus amados pais Carlos e Andrea, à minha irmã Caroline, à minha namorada Ayeska e aos meus colegas de moradia Gustavo e Diego. Agradeço principalmente à minha falecida avó Maria Helena, heroína que me deu apoio, incentivo nas horas difíceis, de desânimo e cansaço.

Meus agradecimentos aos amigos Diego Carvalho, Diogo Santana, Pedro Souza, Antônio Gonçalves, Luiz Felipe, Lucas Félix, Millas Násser, Leandro Campos, Massilon Lourenço, companheiros de trabalhos e irmãos na amizade que fizeram parte da minha formação e que vão continuar presentes em minha vida.

Ao Professor Leonardo Rocha. Companheiro de caminhada ao longo do Curso de Ciência da Computação. Eu posso dizer que a minha formação, inclusive pessoal, não teria sido a mesma sem a sua pessoa. Aos meus colegas Washington Luiz e Felipe Viegas, que foram tão importantes na minha vida acadêmica e no desenvolvimento desta monografia.

Ao Curso de Ciência da Computação da UFSJ, e às pessoas com quem convivi nesses espaços ao longo desses anos. A experiência de uma produção compartilhada na comunhão com amigos nesses espaços foram a melhor experiência da minha formação acadêmica. A todos aqueles que de alguma forma estiveram tão próximos de mim, fazendo esta vida cada vez mais a pena.

"Nada existe.
Tudo é um sonho.
Deus, o homem, o mundo, o Sol e a Lua, a
imensidão das estrelas, um sonho, tudo um sonho.
Coisas que não existem.
Nada existe a não ser o espaço vazio... e você!
E você não é você. Não tem corpo, nem sangue, nem
ossos, você é apenas um pensamento."
(Mark Twain, O Estranho Misterioso)

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. Especificamente, nossa estratégia, chamada CluWords, explora as palavras mais próximas de um determinado espaço *word embedding* pré-treinado para gerar meta palavras capazes de melhorar a representação de documentos, em termos de informações sintáticas e semânticas. Nossa solução inclui as contribuições: (i) a introdução de uma nova representação de dados para modelagem de tópicos baseada em relações sintáticas e semânticas derivadas de distâncias calculadas dentro de um espaço *word embedding* pré-treinado e (ii) a proposta de uma nova estratégia baseada em TF-IDF, especialmente desenvolvida para ponderar as CluWords. Em nossa extensa avaliação experimental, abrangendo 12 bases de dados e 8 linhas de base no estado da arte, apresentamos melhoras (com poucos empates) em quase todos os casos, com ganhos de mais de 50% contra as melhores linhas de base (alcançando até 80% contra alguns). Por fim, mostramos que nosso método é capaz de melhorar a representação de documentos para a tarefa de classificação automática de texto.

Palavras-chaves: Aprendizado de Máquina. Processamento de Linguagem Natural. Modelagem de Tópicos.

Abstract

In this work, 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. Specifically, 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. The novel contributions of our solution include: (i) the introduction of a novel data representation for topic modeling based on syntactic and semantic relationships derived from distances calculated within a pre-trained word embedding space and (ii) the proposal of a new TF-IDF-based strategy, particularly developed to weight the CluWords. In our extensive experimentation evaluation, covering 12 datasets and 8 state-of-the-art baselines, we exceed (with a few ties) in almost cases, with gains of more than 50% against the best baselines (achieving up to 80% against some runner-ups). Finally, we show that our method is able to improve document representation for the task of automatic text classification.

Keywords: Machine Learning. Natural Language Preprocessing. Topic Modelling.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AM	Aprendizado de Máquina
ARTM	Additive Regularization of Topic Models
BOW	Bag-Of-Words
BTM	Biterm Topic Model
CluWords	Cluster of Words
DMM	Dirichlet Multinomial Mixture
DT	Decomposição de Tópicos
ET	Extração dos Tópicos
ETM	Embedding-based Topic Model
ICA	Independent Component Analysis
IDF	Inverse Document Frequency
LDA	Latent Dirichlet Allocation
LF-DMM	Latent Feature with Dirichlet Multinomial Mixture
LSI	Latent Semantic Indexing
LTM	Lifelong Topic Model
NMF	Non-negative Matrix Factorization
NPMI	Métrica Normalized Point-wise Mutual Information
PCA	Principal Component Analysis
PLN	Processamento de Linguagem Natural
PMI	Métrica Point-wise Mutual Information
pLSA	Probabilistic Latent Semantic Analysis
pLSI	Probabilistic Latent Semantic Indexing
RD	Representação de Dados

SeaNMF	Semantics-Assisted Non-negative Matrix Factorization
STE	Skip-gram Topical word Embedding
STOC	Semantic TOpic Combination
SVD	Singular Value Decomposition
TF	Term Frequency
TF-IDF	Term Frequency - Inverse Document Frequency
UTOPIAN	User-driven Topic modeling based on Interactive Non-negative Matrix Factorization
VSMs	Vector Space Models
W2V-L1	Métrica baseada na representação das palavras utilizando a distância L_1

Lista de símbolos

$\mathbb{D}$	Conjunto de documentos
$\mathbb{V}$	Vocabulário do conjunto $\mathbb{D}$
$\mathbb{W}$	Conjunto de vetores, em um espaço *word embedding*, de cada termo em $\mathbb{V}$
Θ	Distribuição de tópicos em um documento em modelos probabilísticos
Φ	Distribuição de termos em um tópico em modelos probabilísticos
ω	Semelhança de cosseno entre dois vetores
Σ	Matriz que captura a força de cada tópico no algoritmo SVD
α	Limiar de similaridade que controla a inclusão do valor de similaridade nas CluWords
μ	Média dos valores dos pesos de uma CluWord
L	Probabilidade de log do conjunto de dados de treinamento, utilizado pelo algoritmo pLSA puro
R	Regularizadores utilizados no ARTM
A	Matriz esparsa representando nas linhas cada documentos e nas colunas cada palavra do vocabulário
C	Matriz de CluWords de uma coleção de documentos
H	Matriz que codifica a relação entre documentos e tópicos no algoritmo NMF
U	Matriz que relaciona termos e tópicos no algoritmo SVD
V	Matriz que captura o relacionamento entre documentos e tópicos no SVD
W	Matriz que codifica a relação entre tópicos e termos no algoritmo NMF
c_{tf-idf}	Métrica TF-IDF Coherence
C_{TF-IDF}	Versão modificada do TF-IDF para as CluWords
C_{TF}	Versão modificada do TF para as CluWords

tf	Frequência de ocorrência de um termo em um documento
idf	Importância de um termo em relação à coleção de documentos
$NPMI$	Métrica Normalized Pairwise Point-wise Mutual Information
tf_idf	Razão entre a frequência de um termo em um documento e a importância do termo na coleção de documentos

Sumário

1 Introdução

1.1 Contexto

Vivenciamos um mundo em que, graças à Internet e suas aplicações/componentes, informações vêm sendo geradas em um ritmo cada vez maior e mais acelerado. Esse grande volume de dados disponível na Web gerou nos últimos anos um diferenciado, desafiante e intrigante cenário para variadas aplicações: há mais dados que efetivamente se pode analisar, como afirmado em (AUDEN, 1996), "Muita informação é tão ruim quanto nenhuma". Representar adequadamente tais informações, sem perdas, bem como desenvolver estratégias efetivas e eficientes para manuseá-las e avaliá-las, é uma das tarefas mais desafiadoras em Ciência da Computação. A Modelagem de Tópicos está entre as abordagens mais exploradas para extrair e organizar informações de grandes quantidades de dados. Essa abordagem visa encontrar tópicos semânticos a partir de documentos textuais (e.g., revisões de produtos e aplicativos, tweets) (BICALHO, 2014). Os tópicos extraídos automaticamente por meio dessas técnicas podem ser explorados por outras aplicações, tais como máquinas de busca e sistemas de recomendação, para auxiliar a realização de tarefas específicas.

O principal objetivo da técnica de Modelagem de Tópicos é descobrir padrões de uso de palavras e conectar documentos que compartilham padrões similares. Assim, a ideia da técnica de Modelagem de Tópicos é que um documento é composto por tópicos ou temas, sendo que, um tópico é composto por uma coleção de palavras que o representa como um todo. Em outras palavras, Modelagem de Tópicos é um modelo para padrões de similaridade entre documentos (ALGHAMDI; ALFALQI, 2015). Portanto, essa técnica é uma tarefa em Aprendizado de Máquina (AM) que extrai tópicos "implícitos" de uma coleção de documentos e atribui os mais prováveis para cada documento (BLEI, 2012).

A Modelagem de Tópicos é uma importante área de pesquisa, principalmente quando não há taxonomia explícita, ou não há um esquema de classificação para associar com documentos, ou quando a rotulagem na classificação é muito incômoda e/ou onerosa de se obter. Essa área se demonstrou importante nos seguintes cenários: (i) expansão da representação para documentos curtos, um problema muito comum em aplicativos de computação social (ZHAO, 2011; PEDROSA, 2016; JIN, 2011); (ii) tarefas de aprendizado não supervisionado (e.g., *clustering*) (XIE; XING, 2013); ou (iii) tarefas de aprendizado supervisionado (e.g., classificação de tópicos) (RUBIN, 2012).

Nesse contexto, representar apropriadamente os dados disponíveis, sem perder informações úteis, assim como desenvolver estratégias efetivas e eficientes para tratar,

organizar e avaliar dados relevantes, são primordiais. As técnicas de Modelagem de Tópicos estão entre as estratégias mais exploradas para extrair e organizar as informações de grandes quantidades de dados. Entretanto, as estratégias tradicionais para identificação de tópicos enfrentam um grande desafio: a semântica dos tópicos é geralmente extraída apenas analisando as propriedades sintáticas dos dados textuais.

Basicamente, o processo de Modelagem de Tópicos pode ser dividido em três blocos de construção principais: (i) **Representação de Dados**, (ii) **Decomposição dos Tópicos**, e (iii) **Extração dos Tópicos**. Resumidamente, a Representação de Dados tem relação com a forma de se codificar os dados textuais (documentos) em uma estrutura que seja capaz de representar as informações relevantes de maneira adequada. Atualmente, a estratégia mais comumente utilizada é a *Bag-of-Words* (BOW) (SALTON; BUCKLEY, 1988), onde a coleção de textos é representada por uma matriz onde cada linha corresponde a um documento, as colunas representam o vocabulário da coleção. Cada posição da matriz representa a importância de uma determinada palavra para um determinado documento (e.g., TF, IDF, TF-IDF). A Decomposição de Tópicos corresponde a tarefa de decompor a matriz que representa uma determinada coleção em outras matrizes que capturam as relações latentes entre documentos e palavras. Existem métodos diversos para esse propósito, desde estratégias probabilísticas tais como o *Latent Dirichlet Allocation* (LDA) (BLEI, 2003) e *Latent Semantic Indexing* (LSI) (DEERWESTER, 1990), até abordagens não probabilísticas, baseadas em álgebra linear, tal como *Principal Component Analysis*(PCA) (PEARSON, 1901) e *Non-negative Matrix Factorization* (NMF) (LEE; SEUNG, 1999). Finalmente, a etapa de Extração de Tópicos visa analisar as matrizes fatoradas na etapa anterior e, por meio das relações latentes entre documentos e palavras, extrair os tópicos que emergem a partir dessas relações. O grande desafio associado a estratégias tradicionais de Modelagem de Tópicos é que a semântica dos tópicos (etapas ii e iii) é geralmente extraída analisando apenas as propriedades sintáticas dos dados textuais (etapa i), assumindo uma alta correlação entre sintaxe e semântica.

Assim, mesmo com as estratégias existentes atualmente para representarem os dados textuais, ainda há informações de contexto que são perdidas na codificação de um documento, fazendo com que a qualidade dos tópicos seja afetada de forma negativa.

1.2 Objetivo

O objetivo deste trabalho é aprimorar os tópicos gerados na Modelagem de Tópicos, sendo que, para atingir tal objetivo, é utilizado modelos não probabilísticos e representações de dados semanticamente enriquecidas baseadas em *word embeddings* (MIKOLOV, 2013). Para que esse objetivo maior seja alcançado, traçamos como objetivos menores:

1. Realizar uma análise das representações de dados atuais baseadas em *word embed-*

dings, tais como Word2Vec (MIKOLOV, 2013), GloVe (PENNINGTON, 2014) e FastText (MIKOLOV, 2017);

2. Avaliar e realizar as adaptações necessárias para que cada uma das etapas do processo de Modelagem de Tópicos seja capaz de utilizar a nova representação proposta;

3. Efetuar a mensuração da qualidade da nova estratégia em bases de dados populares, comparando os resultados obtidos frente as técnicas consideradas estado da arte.

1.3 Justificativa

Word embedding consiste na combinação de métodos de Processamento de Linguagem Natural (PLN) e AM, com o objetivo de mapear as palavras de uma determinada coleção em vetores numéricos, gerados a partir de análises semânticas de palavras extraídas de diversas e enormes coleções textuais, oriundas de diversos contextos. Estes vetores são denominados *Vector Space Models* (VSMs), representando palavras em um espaço vetorial contínuo, assumindo que palavras com significados semânticos similares são vetorialmente próximas. A técnica *word embedding* geralmente traz representações mais ricas que, em última análise, aprimoram os recursos para aprendizado de modelos em diversas áreas, tais como classificação automática de documentos (TANG, 2015) e análise de sentimento (ROTHE, 2016). Com relação ao processo de Modelagem de Tópicos, encontra-se algumas propostas baseadas em métodos probabilísticos que utilizam a riqueza provida por representações *word embeddings* (NGUYEN, 2015; LI, 2017; DAS, 2015). Nesse caso, a técnica *word embedding* é utilizada de forma auxiliar, apenas no processo de ajustes dos modelos probabilísticos.

Apesar de modelos não probabilísticos serem de propósito mais geral e, normalmente, apresentarem resultados melhores, a principal razão para a ausência de mais trabalhos utilizando-os é que o uso das informações mais ricas fornecida pela representação *word embeddings* dificulta a Modelagem de Tópicos. Há uma falta de correspondência direta entre tópicos e unidades semânticas menores (e.g., palavras) nessas representações mais ricas. O que justifica nosso objetivo geral, bem como os objetivos específicos (1) e (2). Com relação ao objetivo específico (3), é feito uma comparação dos resultados obtidos em comparação às técnicas existentes no estado da arte na área de Modelagem de Tópicos. Para comparação da nossa técnica com as existentes no estado da arte, foram utilizadas 12 bases de dados que são popularmente utilizadas em PLN na área de AM.

1.4 Contribuições

Em resumo, nossas principais contribuições neste trabalho são:

- Uma nova representação de documentos (CluWords) que explora, em um framework unificado, relações semânticas e sintáticas entre palavras em uma coleção de documentos com o objetivo de aprimorar a Modelagem de Tópicos não probabilística;
- A proposta de uma nova estratégia de ponderação do TF-IDF para as CluWords;
- Uma extensa avaliação experimental abrangendo 12 bases de dados e 8 linhas de base no estado da arte, em que nossa abordagem se destaca (com alguns empates) em quase todos casos, com ganhos de mais de 50% contra as melhores linhas de base (alcançando até 80% contra algumas).

1.5 Organização do Documento

Esse trabalho é organizado da seguinte maneira. No Capítulo 2 apresentamos uma extensa avaliação sobre as técnicas existentes em Modelagem de Tópicos, apresentando com mais destaque as que serão utilizadas como linhas de base. No Capítulo 3 é detalhado a estratégia de CluWords proposta. No Capítulo 4 introduzimos nossa avaliação experimental. Por fim, no Capítulo 5 conclui-se o trabalho e discutimos os trabalhos futuros.

2 Trabalhos Relacionados

Neste capítulo, apresentamos alguns trabalhos, tradicionais e recentes, relevantes na área de modelagem de tópicos e discutimos seus pontos positivos e negativos, bem como suas diferenças. Os trabalhos são apresentados de acordo com os três blocos definidos na introdução do trabalho: Representação de Dados, Decomposição de Tópicos e Extração dos Tópicos. Primeiramente, mostramos as técnicas utilizadas para Representação de Dados. Nesse caso, mostramos os algoritmos tradicionais na área de PLN e as representações de *word embeddings* que são as técnicas estado da arte atualmente. Depois, introduzimos as técnicas de Decomposição de Tópicos, mostrando as variações que são baseadas em modelos probabilísticos e as que se utilizam de modelos não probabilísticos. Por fim, são discutidas as técnicas de extração de tópicos, que visam a identificar os tópicos relevantes depois que os tópicos latentes são descobertos. Em cada seção, também é mostrado como a técnica proposta por este trabalho difere dos trabalhos relacionados.

2.1 Representação de Dados

Seja uma base de dados textual (i.e., coleção) representada por $\mathbb{D} = \{d_1, ..., d_n\}$, onde cada elemento do conjunto é um documento e considerando que o vocabulário (todas palavras contidas na base) é representado por $\mathbb{V} = \{w_1, ..., w_m\}$. Cada documento em $\mathbb{D}$ é representado por um conjunto de palavras de $\mathbb{V}$, sendo que, geralmente $\mathbb{V}_d \ll \mathbb{V}$. Neste trabalho a notação "termo" é utilizada tanto para referenciar uma palavra ou um conjunto de palavras. Assim, a etapa de representação de dados consiste na representação de cada termo $w \in \mathbb{V}$ em cada documento $d \in \mathbb{D}$, onde é necessário extrair informações úteis de cada termo. O principal objetivo da representação de dados em um sistema de ponderação de termo é o aprimoramento da eficácia das informações recuperadas de cada termo.

A forma de representação de um texto mais conhecida, e também uma das mais utilizadas, é a *Term Frequency - Inverse Document Frequency* e suas variações (TF-IDF) (SALTON; BUCKLEY, 1988). Essa estratégia consiste no cálculo do valor TF-IDF de cada palavra em cada documento de uma base de dados. O fator TF captura de um termo w o quão frequente ele é em um documento, enquanto o fator IDF identifica a importância desse termo em relação a todos documentos. A representação por TF-IDF gera um vetor esparso $|\mathbb{V}|$-dimensional onde cada termo tem seu valor de TF-IDF representado em cada documento (com entradas diferentes de zero correspondendo ao conjunto de termos $\mathbb{V}_d \subset \mathbb{V}$ observado nos documentos).

Embora a estratégia TF-IDF seja, de longe, a mais usada (especialmente considerando as abordagens de aprendizagem baseadas em modelos de espaço vetorial), a

estratégia não leva em consideração a importância da coocorrência de palavras em um documento. Em determinados contextos, a coocorrência de palavras pode ser bastante significativa para a eficácia do aprendizado de um modelo. Uma estratégia simples para superar isso é usar o método *n-grams* (CAVNAR, 1994). O método *n-gram* considera uma sequência de n palavras que coocorrem; ou seja, para uma palavra w o método considera n palavras antes e/ou depois para representar um novo termo. Desse modo, o mesmo valor de TF-IDF é usado, só que aplicado às *n-grams*. O uso de *n-grams* já demonstrou melhorias significativas em vários domínios de PLN (e.g., classificação de documentos, compreensão de linguagem natural, etc.), apesar de limitar a captura de informações contextuais observadas em padrões não sequenciais.

Recentemente, muito tem sido desenvolvido em termos de melhorar a representação de dados. Ultimamente, a técnica com resultados mais promissores são os modelos de *word embedding*, como Word2Vec(MIKOLOV, 2013), GloVe(PENNINGTON, 2014) e FastText (MIKOLOV, 2017) que, baseado nas estatísticas de coocorrência de palavras em bases de dados de texto, representam palavras em vetores tais que suas semelhanças se correlacionam com a relação semântica (e.g., termos adjacentes de um termo alvo). Para este fim, essas estratégias fazem o uso da informação contextual, como os termos adjacentes a outro termo. Como mostrado em (BARONI, 2014), os modelos de previsão superam consistentemente os modelos de contagem em várias tarefas, como categorização de conceito, detecção de sinônimos e tarefas de relacionamento semântico, fornecendo fortes evidências em favor da suposta superioridade dos modelos de *word embedding*.

Com objetivo de obter uma representação mais rica dos dados e com menor custo computacional, Mikolov, Chen, Corrado e Dean (2013) propuseram o modelo chamado Word2Vec – provavelmente a estratégia de *word embeddings* mais popular atualmente. A estratégia não supervisionada visa aprender vetores de alta qualidade para cada palavra existente no modelo treinado. Isso é obtido por uma rede neural treinada com sequências de palavras que coocorrem dentro de uma janela de tamanho fixo, a fim de prever a enésima palavra, dada as palavras $[w_1, ..., w_{n-1}]$ ou o contrário. Ou seja, cada palavra é utilizada como entrada de um classificador log-linear com uma camada de projeção contínua, onde, o modelo prediz palavras dentro de uma certa distância (antes ou depois) da palavra corrente. Uma vez que palavras mais distantes são geralmente menos relacionadas à palavra atual do que àquelas próximas a ela, é colocado um peso de menor importância às palavras distantes, analisando com menos importância essas palavras no modelo utilizado. A saída é uma matriz de vetores de palavras representadas em um espaço vetorial. O método é capaz de capturar as relações semânticas entre palavras dado um grande conjunto de dados textuais de maneira eficiente e eficaz, sendo que, o modelo também possibilita operações aritméticas entre palavras. Assim, por exemplo, se for feito as operações $vetor(\text{“}King\text{”}) - vetor(\text{“}Man\text{”}) + vetor(\text{“}Woman\text{”})$ o resultado será um vetor próximo do vetor de representação da palavra *Queen*. Diferente de outros modelos

de distribuição, tanto o Word2Vec quanto o GloVe são modelos de previsão, no sentido de que visam prever a ocorrência de palavras ao invés de depender apenas de padrões de coocorrência para representação dos dados. Isso geralmente traz representações mais ricas que, em última análise, melhoram as capacidades de aprendizado dos modelos.

Os modelos baseados em Word2Vec têm sido usados para capturar informações semânticas e sintáticas usando aritmética vetorial, mas a origem dessas informações permanece opaca no procedimento. Pennington, Socher e Manning (2014) apresentam o modelo GloVe, que utiliza das técnicas de contagem de palavras (TF, TFIDF) enquanto capturam as subestruturas lineares predominantes nos métodos baseados em previsão *log-bilinear*, como o Word2Vec. Este modelo, aproveita eficientemente das informações estatísticas, treinando apenas apenas nos elementos não nulos em uma matriz de coocorrência de palavras (i.e., *n-gram*), ao invés de toda a matriz esparsa. O modelo produz um espaço vetorial com subestrutura significativa, como evidenciado pelo desempenho de 74% em uma tarefa de analogia de palavras. O resultado é um modelo de regressão *log-bilinear* global para aprendizagem não supervisionada de representações de palavras e tarefas de reconhecimento de entidades nomeadas (sub tarefa de extração de informação que visa localizar e classificar as menções de entidades).

O FastText (MIKOLOV, 2017), por outro lado, aprende vetores para as sub palavras (e.g., termos *n-grams*) encontrados dentro de cada palavra, bem como a palavra completa. Em cada etapa de treinamento no FastText, a média da palavra alvo e os vetores de sub palavras são usados para treinamento. O ajuste calculado a partir do erro é então usado uniformemente para atualizar cada um dos vetores que foram combinados para formar o alvo. Isso adiciona muito custo de computação à etapa de treinamento. O *trade-off* é um conjunto de vetores de palavras que contém informações de sub palavras embutidas. Os autores afirmam que os potenciais benefícios do FastText são: (i) gera melhor incorporação de palavras para palavras raras; (ii) O uso de incorporação de caracteres para tarefas de recebimento de dados aumentou o desempenho dessas tarefas em comparação ao uso de incorporação de palavras como Word2Vec e GloVe.

Outra forma comum de explorar a técnica de *word embeddings* para representação de documentos é conhecida como *Paragraph2Vec* (LE; MIKOLOV, 2014). O *Paragraph2Vec* estende um modelo de rede neural que representa palavras (MIKOLOV, 2013) em sentenças, parágrafos ou documentos de tamanho arbitrário. A principal intuição por trás do *Paragraph2Vec* é a construção de vetores de documentos densos treinados para prever palavras para cada documento. O *Paragraph2Vec* aprende vetores de parágrafos em vez de vetores de palavras, sendo esse método inspirado pelo aprendizado de vetores de palavras. Assim sendo, o *Paragraph2Vec* faz a aprendizagem dos vetores de cada palavra individual para depois ser feito a aprendizagem do vetor que representa parágrafos em uma coleção (já que um parágrafo é um aglomerado de palavras). Em contraste com

o *Paragraph2Vec*, outra maneira de representar documentos usando *word embeddings* é pela técnica de *Pooling* (LEV, 2015). Esta estratégia defende que, agrupando adequadamente vetores de palavras em uma sentença, pode haver resultados pelo menos muito competitivos. Sendo assim, a abordagem de *Pooling* representa uma sentença como um multiconjunto de vetores Word2Vec. É afirmado que a notação de multiconjunto é utilizada para clarificar que a ordem das palavras em uma sentença não afeta a representação final desta, sendo que um vetor pode aparecer mais de uma vez. Essa suposição é baseada nos benefícios do uso de métodos de *Pooling* como uma das principais etapas em muitos domínios de visão computacional antes que as técnicas de aprendizado profundo começassem a desfrutar do desempenho de última geração. Os autores avaliaram o uso da técnica PCA (SÁNCHEZ, 2013) e *Independent Component Analysis* (ICA) (HYVÄRINEN; OJA, 2000) como uma etapa de preprocessamento não supervisionada que transforma o espaço vetorial semântico em canais semânticos independentes. Após a etapa de preprocessamento, os autores propõem o uso de *Fisher Vectors* (PERRONNIN; DANCE, 2007) como uma estratégia de *Pooling* para incorporação de palavras, uma vez que é a atual técnica de *Pooling* no estado da arte usada em diferentes aplicações (CHATFIELD, 2011).

Diferente dos métodos mencionados anteriormente, é proposto neste trabalho uma nova técnica para aprimorar a representação de documentos, que foi nomeada *Cluster of Words* (CluWords), que se aproveita das informações dos modelos de *word embeddings*. Especificamente, a estratégia criada por este trabalho explora as palavras mais próximas de um determinado modelo *word embedding* para gerar "meta palavras" capazes de melhorar a representação de um documento, em termos de informação sintática e semântica. A exploração explícita da similaridade entre *word embeddings* para encontrar as palavras mais próximas fornece informações refinadas sobre as relações entre as palavras. A estratégia CluWords combina as evidências sintáticas tradicionais (i.e., ocorrência de palavras em um documento) e a similaridade entre uma palavra e suas vizinhas. Portanto, espera-se suavizar as possíveis desvantagens de usar o espaço projetado de *word embeddings*, sendo a principal desvantagem a grande quantidade de informação sobre todo vocabulário, explorando apenas evidências claras de similaridade e confiando nas representações sintáticas tradicionais de documentos. Cada palavra do vocabulário corresponde a uma CluWord, que é ponderada de acordo com a nova estratégia de ponderação baseada no TF-IDF, particularmente desenvolvida para medir a importância de uma determinada CluWord para definir um tópico de um documento. Essa nova representação é rica e flexível o suficiente para ser explorada por qualquer tipo de abordagem de modelagem de tópicos.

2.2 Decomposição dos Tópicos

Nessa seção são detalhados os algoritmos que visam a descoberta de tópicos abstratos a partir da representação dos dados de uma coleção de documentos. Primeiramente, são

discutidos os modelos não probabilísticos. Nesse caso, os dados são modelados de acordo com os dois principais conceitos a seguir: (i) cada documento segue uma distribuição de tópico $\Theta^{(d)}$ e (ii) cada tópico segue uma distribuição de termos $\Phi^{(z)}$.

Seja $\Theta^{(d)} = P_d(z)$ a distribuição do tópico sobre os tópicos z observados em um documento d. Além disso, seja $\Phi^{(z)} = P(w|z)$ a distribuição de probabilidade sobre os termos w considerando documentos pertencentes ao tópico abstrato z. $\Theta^{(d)}$ e $\Phi^{(z)}$ refletem em que medida um termo é importante para um tópico e em que medida um tópico é importante para um documento, respectivamente. Ambos os conceitos são fundamentais para o chamado algoritmo *Probabilistic Latent Semantic Indexing* (pLSI) (HOFMANN, 1999), que leva em conta a probabilidade de coocorrência de termos e documentos $P(d, w)$ como uma combinação multinomial de distribuições condicionalmente independentes. Mais especificamente, seja $P(z|d)$ a probabilidade condicional de observar um tópico z dado um documento d, baseado em $\Theta^{(d)}$, e $P(w|z)$ ser a probabilidade condicional de observar um termo w dado um tópico z. Assim, $P(d, w) = P(d)P(w|d)$ e, marginalizando, $P(w|d) = \sum_z P(w|z)P(z|d)$.

Observe que o pLSI não leva em conta como $P(w|z)$ (Φ) é gerado (e, em última análise, também $P(z)$), sendo mais difícil generalizar para cenários do mundo real. Por outro lado, em (DAS, 2015) os autores propõem a assim chamada *Latent Dirichlet Allocation* (LDA), que generaliza o pLSI em termos de como $P(w|z)$ é estimado: ele assume uma distribuição de *Dirichlet* – uma distribuição multivariada contínua que reflete o fato de que os documentos geralmente contribuem para apenas alguns tópicos e tópicos geralmente contribuem para um pequeno conjunto de palavras. Apesar de ser uma das estratégias mais utilizadas para a decomposição de tópicos latentes, o LDA tem seus próprios desafios, como a escassez de dados e a geração de tópicos incoerentes. Em (CHENG, 2014), os autores propuseram o método de *Biterm Topic Model* (BTM) para lidar com o desafio de dispersão dos dados, através do uso do que eles chamam de termos "bi-gerados" com base em estatísticas de coocorrência de termos frequentes. Especificamente, os autores do BTM consideram que todo o *corpus* é uma mistura de tópicos, onde cada bitermo é desenhado para um tópico específico de forma independente. A probabilidade de que um bitermo pertença à um tópico é capturado pela chance de ambas palavras no bitermo seja pertencente ao tópico. Com essa abordagem, consegue-se modelar diretamente o padrão de coocorrência entre palavras, em vez de uma única palavra, como uma unidade que transmite a semântica dos tópicos. Sem dúvida, a coocorrência de um par de palavras pode revelar muito melhor os tópicos do que a ocorrência de uma única palavra e, então, aprimorar o aprendizado dos tópicos. Chen e Liu (2014) lidam com tópicos incoerentes por meio 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 informações para produzir tópicos mais coerentes e confiáveis. A suposição básica aqui é que as relações léxicas e semânticas são fundamentais para descobrir

tópicos coerentes.

No pLSA puro (HOFMANN, 1999), as distribuições palavra-tópico (Φ) e as distribuições documento-tópico (Θ) são aprendidas otimizando diretamente a probabilidade de log do conjunto de dados de treinamento $L(\Phi, \Theta)$. Recentemente, Vorontsov e Potapenko (2015) desenvolveram a abordagem de *Additive Regularization of Topic Models* (ARTM), o modelo básico de pLSA 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 2.1.

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R(\Phi, \Theta), \;\; L(\Phi, \Theta) + R(\Phi, \Theta) \rightarrow \max(\Phi, \Theta) \tag{2.1}$$

Embedding-based Topic Model (ETM) (QIANG, 2017) é outra técnica que incorpora o conhecimento externo de correlação de palavras em textos curtos para melhorar a coerência da modelagem de tópicos. O ETM não só resolve o problema das 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 melhor chance de serem colocadas no mesmo tópico. Os métodos LDA, BTM, LTM, ARTM e ETM são usados como linhas de base neste trabalho para comparação de resultados.

Guzman e Maalej (2014) propuseram o método FS, que é uma estratégia usada para construir tópicos com informações de sentimento. Ele extrai palavras que coocorrem frequentemente (também conhecido como *bi-grams*). Em seguida, infere a força do sentimento dos *bi-grams* extraídos com base na pontuação de sentimento dos documentos em que ocorreram. Para gerar os tópicos, a estratégia aplica o LDA a esses *bi-grams* sentimentais. O método FS também é utilizado para comparação de resultados obtidos.

Considera-se agora as técnicas modelagem de tópicos não probabilísticas, compreendendo estratégias como a fatoração de matriz. Nesse caso, 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 modelagem de tópicos, é a *Singular Value Decomposition* (SVD) (GOLUB; REINSCH, 1970). Nesse caso, a matriz A é fatorada em três matrizes, a saber, $U_{n,k}$, $\Sigma_{k,k}$ e $V_{m,k}$, tal que $A = U\Sigma V^t$. A matriz U captura a relação entre termos e tópicos, em que V captura o relacionamento entre documentos e tópicos. A matriz Σ captura a força de cada tópico nos dados de entrada. Sob este novo espaço semântico, produzido pelas submatrizes, termos e documentos

similares são geralmente localizados em regiões vizinhas, de acordo com alguma métrica de distância ou similaridade. Finalmente, outra estratégia amplamente utilizada para a decomposição de matrizes aplicável à modelagem de tópicos é a *Non-negative Matrix Factorization* (NMF) (LEE; SEUNG, 1999). 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. Observe que, ao contrário do NMF, os fatores SVD podem conter entradas positivas e negativas. Ao lidar com dados textuais adequadamente representados, a matriz A geralmente contém pontuações de termo não negativas, como TF-IDF, com semântica bem definida (por exemplo, frequência de termo e raridade). É natural esperar que os fatores extraídos sejam não negativos, de modo que tal semântica possa ser de alguma forma preservada. O NMF é usado nesse trabalho como estratégia de fatoração de matriz para a modelagem de tópicos.

Trabalhos recentes têm sido propostos para melhorar a construção de tópicos por meio do uso de *word embeddings* como uma informação auxiliar para a modelagem probabilística de tópicos. Nguyen, Billingsley, Du e Johnson (2018) propuseram um modelo de tópico chamado *Latent Feature with Dirichlet Multinomial Mixture* (LF-DMM), sendo esse uma combinação de fatores latentes (i.e., vetores de palavras pré treinados) e o modelo DMM para textos curtos (onde cada documento assume pertencer a somente um tópico). Esse método substitui a distribuição de palavras de um tópico por uma mistura de dois componentes: um componente multinomial de Dirichlet e um componente de incorporação de palavras contínuo. Dessa forma, o modelo pode se beneficiar de informações de relacionamento semântico entre palavras para gerar tópicos melhores. Das, Zaheer e Dyer (2015) propõem um modelo de tópico baseado em LDA usando a distribuição gaussiana multivariada com *word embeddings*.

Em (SHI, 2017), os autores propõem o *framework Skip-gram Topical word Embedding* (STE), que pode aprender a incorporar vetores de palavras gerados pelo método *word embeddings* e tópicos latentes de maneira unificada. É mostrado como é possível reunir os conceitos citados para explorar o reforço mútuo entre eles, mostrando que as estratégias utilizadas são complementares se utilizadas como um processo único. Li *et al.* (LI, 2017) propõem 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. O GPU-DMM é outra linha de base utilizada para comparação de resultados. Finalmente, em (SHI, 2018) os autores propõem um modelo de fatoração da matriz resultante do método NMF auxiliado pela semântica – *Semantics-Assisted Non-negative Matrix Fac-*

torization (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*, sendo que, essa estratégia se mostrou eficaz para revelar as relações semânticas das palavras. O SeaNMF é outro método utilizado neste trabalho como linha de base.

Até onde sabe-se, não há trabalho que combine as informações dos vetores das palavras em modelos de *word embeddings* e modelos não probabilísticos. A principal razão para a ausência de trabalhos combinando tais informações é que a introdução da informação mais rica fornecida pela representação de *word embeddings* de palavras dificulta a representação de tópicos devido à falta de correspondência direta entre tópicos e unidades semânticas menores (e.g., palavras) nessas representações mais ricas.

2.3 Extração dos Tópicos

Depois que os tópicos abstratos (latentes) são descobertos, é importante finalmente identificar os relevantes. Como mencionado anteriormente, as estratégias probabilísticas e não probabilísticas para identificação de tópicos tendem a gerar tópicos incoerentes e/ou irrelevantes, o que é indesejável. Nesta subseção, são mostradas as técnicas existentes para identificar os tópicos relevantes nos tópicos gerados previamente e tratá-los de forma adequada.

Os autores em (CHOO, 2013) propõem uma estratégia supervisionada chamada UTOPIAN. Essa estratégia se baseia em julgamentos humanos para revisar os tópicos identificados. O método provê uma grande variedade de interações capazes de melhorar a modelagem de tópicos. Ele permite que agentes humanos filtrem tópicos incoerentes, mesclem tópicos distintos, excluam termos de tópicos identificados e criem novos tópicos. Essas operações são realizadas com o auxílio de uma interface gráfica para o usuário. Para um processo totalmente automático, os autores em (DEERWESTER, 1990) propõem o uso do LSI seguido por um algoritmo de agrupamento para identificar tópicos semânticos relevantes. Mais especificamente, os k fatores latentes são identificados a partir de dados brutos, que podem ser representados no novo espaço semântico k-dimensional. Os pontos são então agrupados usando uma matriz de covariância representada no topo deste novo espaço semântico para descobrir os tópicos relevantes.

Recentemente, em (BICALHO, 2014) uma nova técnica, baseada em NMF, foi proposta: o método denominado *Semantic TOpic Combination* (STOC). O objetivo principal dessa técnica é gerar tópicos semânticos mais coesos e significativos dentro de um contexto. Resumidamente, após a fatoração NMF da matriz A, um grafo tripartite ponderado é construído sobre matrizes H e W (obtidas da fatoração NMF, $A \approx H \times W$) ,

onde os pesos vêm dessas submatrizes. Esse gráfico captura a relação entre documentos e termos, bem como termos e tópicos. Um gráfico de transição de tópico é construído através de um passeio aleatório neste gráfico tripartite. Tal procedimento de caminhada aleatória é capaz de descobrir relacionamentos indiretos entre os tópicos, agrupando, em última instância, os semelhantes, por meio da métrica de distância Bhattacharyya. Em cada interação, os dois tópicos com a menor distância são mesclados, até que todos os tópicos sejam finalmente mesclados.

2.4 Considerações Finais

Neste capítulo, apresentamos os principais trabalhos relacionados de acordo com as etapas de Modelagem de Tópicos. Começamos mostrando as técnicas de Representação de Dados, onde o método mais conhecido para extrair dados de documentos é o TF-IDF (SALTON; BUCKLEY, 1988). Foi apresentado as desvantagens que o TF-IDF traz em alguns contextos e como as técnicas de *n-grams* (CAVNAR, 1994), e as baseadas em *word embeddings*(MIKOLOV, 2013; MIKOLOV, 2017; MIKOLOV, 2013; LEV, 2015; SÁNCHEZ, 2013; HYVÄRINEN; OJA, 2000; PERRONNIN; DANCE, 2007; CHATFIELD, 2011), superam essas desvantagens. Realizamos uma comparação dos diferentes modelos de *word embeddings* utilizados atualmente e como a técnica de CluWords se inspira nesses modelos. No final da seção de representação de dados, mostramos qual a ideia por trás da técnica denominada CluWords que propomos neste trabalho.

Em seguida, discutimos sobre as técnicas existentes em Decomposição de Tópicos. Comparamos tanto técnicas probabilísticas quanto técnicas não probabilísticas, apresentando suas vantagens e desvantagens. Começamos com as técnicas probabilísticas, onde, apesar das variações dos algoritmos, eles baseiam-se em dois conceitos: (i) cada documento segue uma distribuição de tópico $\Theta^{(d)}$ e (ii) cada tópico segue uma distribuição de termos $\Phi^{(z)}$. Entre os algoritmos probabilísticos que selecionamos estão: pLSI (HOFMANN, 1999), LDA (DAS, 2015), BTM (CHENG, 2014), LTM (CHEN; LIU, 2014), pLSA (HOFMANN, 1999), ARTM (VORONTSOV; POTAPENKO, 2015), ETM (QIANG, 2017), FS (GUZMAN; MAALEJ, 2014). Em relação às técnicas de modelagem de tópicos não probabilísticas, o objetivo é decompor uma matriz esparsa $A \in \mathbb{R}^{nxm}$, que é uma codificação do conjunto de documentos $\mathbb{D}$ e o seu respectivo vocabulário $\mathbb{V}$, em submatrizes que preservam alguma propriedade ou restrição desejada. Os algoritmos não probabilísticos cobertos neste trabalho são: SVD (GOLUB; REINSCH, 1970), NMF (LEE; SEUNG, 1999), LF-DMM (NGUYEN, 2018), STE (SHI, 2017), GPU-DMM (LI, 2017), SeaNMF (SHI, 2018). Entre a grande diversidade das técnicas apresentadas, selecionamos as seguintes como linhas de base em nossa avaliação experimental reportada no Capítulo 4: LDA, BTM, LTM, ARTM, ETM, FS, GPU-DMM, SeaNMF. A seleção dessas técnicas se deu devido aos resultados reportados na literatura sobre seus desempenhos e por serem

estado da arte em Modelagem de Tópicos.

Finalizando, mostramos as técnicas existentes que são mais utilizadas para Extração de Tópicos. O propósito da extração de tópicos é identificar os tópicos relevantes a partir dos tópicos abstratos (latentes) descobertos. É comparado uma estratégia que iterativa que depende de julgamentos humanos (UTOPIAN) com as estratégias que realizam a descobertas dos tópicos de forma automática (LSI+Algoritmo de Agrupamento (DEERWESTER, 1990), STOC (BICALHO, 2014)). Neste trabalho não utilizamos esta parte da modelagem de tópicos, já que é considerada um complemento para melhorar a qualidade dos tópicos latentes a partir dos resultados do bloco de Decomposição de Tópicos. Entretanto, nosso *framework* possui a capacidade de implementar as técicas de Extração de Tópicos para melhorar a aprimorar os tópicos gerados.

3 Estratégia CluWords

Neste capítulo, descrevemos nossa estratégia para transformar a representação tradicional de *Bag-Of-Words* (BOW) de documentos para incluir informações semânticas relacionadas aos termos presentes nos documentos. O contexto semântico é obtido por meio de um modelo de representação de palavras pré treinado, como Word2Vec (MIKOLOV, 2013) e FastText (MIKOLOV, 2017). Nossa abordagem consiste em transformar cada documento em uma nova representação, na qual as palavras originais são substituídas por um grupo de palavras às quais nos referimos como *Cluster of Words* (CluWords). O processo de transformação é composto de duas fases. Na primeira, computamos, para cada termo w do vocabulário, seu CluWord correspondente. Uma CluWord para um termo w é um conjunto de termos no vocabulário que os vetores de palavras são mais semelhantes ao termo w. Na segunda fase, calculamos uma versão modificada do esquema de ponderação TF-IDF para as CluWords com intenção de explorar esses novos atributos como uma representação mais rica de documentos da coleção. A primeira fase da nossa abordagem é descrita na seção 3.1, enquanto a segunda é apresentada na seção 3.2

3.1 Geração das CluWords

Considerando o vocabulário $\mathbb{V}$ dos termos presentes no conjunto de documentos . Além disso, seja $\mathbb{W}$ o conjunto de vetores representando cada termo em $\mathbb{V}$ de acordo com o modelo *word embedding* pré treinado (e.g., Word2Vec, FastText, etc.). Assim, cada termo w em $\mathbb{V}$ tem um vetor correspondente em $\mathbb{W}$ e cada vetor $u \in \mathbb{W}$ tem comprimento l, onde l é a dimensionalidade do espaço vetorial do modelo.

Definimos as CluWords como uma matriz $C \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$, onde cada índice $C_{w,w'}$ é calculado de acordo com a Equação 3.1

$$C_{w,w'} = \begin{cases} \omega(w,w') & \text{se } \omega(w,w') \geq \alpha \\ 0 & \text{caso contrário} \end{cases} \tag{3.1}$$

onde $\omega(w,w')$ é a semelhança de cosseno definida em 3.2 e α é um limiar de similaridade que controla a inclusão do valor da similaridade entre o termo w e um termo w'. Nesta notação cada CluWord é representado como uma linha C_w e cada coluna w' em $\mathbb{V}$ pode corresponder a um componente em C_w se a semelhança de cosseno entre os vetores w e w' no espaço vetorial de palavras é maior ou igual a um limite α. Caso contrário, a coluna

w' é igual a zero.

$$\omega(u, v) = \frac{\sum_i^l u_i \cdot v_i}{\sqrt{\sum_i^l u_i^2} \cdot \sqrt{\sum_i^l v_i^2}} \tag{3.2}$$

A CluWord C_w relaciona w com suas palavras mais próximas, limitando essa relação com o valor de corte α que filtra as palavras "ruidosas" (i.e., palavras que não tem relação significativa com w) da CluWord. Como o limite α é um valor de similaridade de cosseno, ele está contido no intervalo $[0, 1]$. Se $\alpha = 0$ as semelhanças de todos os termos em $\mathbb{V}_T$ serão incluídas em C_w, caso contrário, se $\alpha = 1$ apenas a similaridade de w a si mesmo (i.e., $\omega(w, w) = 1, 0$) está incluído em C_w. Assim, a seleção apropriada de um valor para o parâmetro α é um aspecto importante da geração de boas CluWords. Além disso, o α controla a dispersão da representação resultante de um documento. Com altos valores de α, existem apenas algumas CluWords relacionadas a um documento. Isso é semelhante à representação BOW tradicional, em que a ocorrência de uma palavra em um documento determina se essa palavra será usada na representação do documento. No entanto, com valores baixos de α, mais CluWords tendem a estar relacionadas ao documento, o que reduz a dispersão da representação do documento. Note que uma vez que selecionamos um valor apropriado para α, cada CluWord C_w mantém os valores das similaridades dos termos mais similares a w de acordo com a semântica estabelecida pelo modelo *word embedding*.

Nossa intenção é usar as CluWords para substituir a representação original pela técnica BOW de documentos. É importante notar que o objetivo das CluWords é (principalmente, mas não apenas) enriquecer a representação BOW, adicionando informação semântica, ou seja, cada termo w será substituído por sua CluWord C_w correspondente em cada documento a que pertence. Para usar a representação CluWord, precisamos calcular o TF-IDF das CluWords. Descrevemos uma forma de calcular a ponderação TF-IDF para as CluWords na Seção 3.2.

Para clarificar como as CluWords extraem informações sintáticas e semânticas de cada palavra, na Tabela 1 apresentamos um exemplo das palavras pertencentes a uma CluWord cujo centroide é a palavra "chat". A tabela mostra as palavras que consideramos, numa análise muito informal, sintaticamente, semanticamente e não relacionadas com o respectivo centroide. A partir desse exemplo, podemos verificar como uma CluWord explora mais informações sobre o contexto que cada palavra está inclusa, além de, capturar as informações sintáticas sobre uma palavra.

Tabela 1 – Exemplo de palavras pertencentes à CluWord que tem como centroide a palavra "chat".

Palavras semanticamente similares	**Palavras sintaticamente similares**	**Palavras não relacionadas**
audio, communicate, communication, contact, conversation, conversations, discuss, email, emails, forum, hear, interact, interaction, listen, listening, mail, message, messages, messaging, news, phone, post, reply, socialize, socializing, speak, talk, talking, voice, meeting, networking, room, service.	chat, chats, chatted, chatting	access, avatar, buddies, buddy, download, dude, evening, evenings, exchange, gallery, game, gaming, girl, girlfriend, guys, homework, interface, mate, mates, pal, server, sip, sit, smilies, strangers, stuff, telephone, thoughts, twitter, video, wander, wanna, web.

3.2 Ponderação TF-IDF para as CluWords

Basicamente, o TF-IDF convencional (SALTON; BUCKLEY, 1988) é uma medida de importância de um termo que avalia dois aspectos distintos: (i) a relevância do termo em um documento específico d (caracterizado pelo componente TF) e (ii) a importância do termo na coleção de documentos $\mathbb{D}$ a serem considerados (dado pelo componente IDF). O componente TF ($tf(w,d)$) contabiliza a frequência de ocorrência do termo w no documento d. O IDF mede a importância do termo w em uma coleção de documentos. Quanto mais um termo ocorre em uma coleção de documentos, menos importante ele é considerado. Assim, o IDF de um termo deve estar inversamente relacionado ao número de documentos em que o termo ocorre. A pontuação TF-IDF $tf\_idf(w,d)$ do termo w no documento d é definido na Equação 3.3.

$$tf\_idf(w,d) = tf(w,d) \cdot log\left(\frac{|\mathbb{D}|}{n_w}\right) \tag{3.3}$$

onde n_w é o número de documentos em $\mathbb{D}$ onde w ocorre.

As CluWords foram criadas com base na semelhança semântica dos termos, de modo que a métrica TF-IDF convencional não é capaz de ponderar esses recursos, aproveitando ao máximo as informações fornecidas por eles. Nossa motivação é combinar os dois aspectos da métrica TF-IDF convencional com a informação semântica de uma CluWord. A seguir, propomos uma versão modificada do TF-IDF para associar as CluWords a serem incluídas na representação BOW estendida do documento d. O TF-IDF de uma CluWord no documento d é definido de acordo com a Equação 3.4.

$$C_{TF-IDF} = C_{TF} \times idf(C) \tag{3.4}$$

Primeiro, o TF ($tf(w,d)$) 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 um termo w no documento d. O TF das CluWords pode ser medido como um produto de matrizes como descrito na Equação 3.5.

$$C_{TF} = C \times T^T \tag{3.5}$$

O valor de $C_{TF_{d,w}}$ corresponde à soma dos produtos das frequências dos termos $T_{d,w'}$ de cada termo $w' \in Cw, w' \neq 0$ ocorrendo no documento d.

Para calcular o IDF de uma CluWord C_w, primeiro definimos o vocabulário $\mathcal{V}_{d,w}$ composto por todos os termos no documento d que têm o peso (ω_w) diferente de zero na CluWord C_w. Isso é formalmente definido na Equação 3.6.

$$\mathcal{V}_{d,C_w} = \{w' \in d | C_{w,w'} \neq 0 \text{ in } C_w\} \tag{3.6}$$

Em seguida, calculamos a média dos valores dos pesos da CluWord C_w dos termos que ocorrem no vocabulário $\mathcal{V}_{d,C_w}$, de acordo com a Equação 3.7.

$$\mu_{C_w,d} = \frac{1}{|\mathcal{V}_{d,C_w}|} \cdot \sum_{w \in \mathcal{V}_{d,C_w}} w_w \tag{3.7}$$

Finalmente calculamos o IDF da CluWord C_w como definido na Equação 3.8, onde $\mathcal{D}$ é o conjunto de treinamento.

$$idf(C_w) = log\left(\frac{|\mathcal{D}|}{\sum_{1\leq d\leq|\mathcal{D}|} \mu_{C_w,d}}\right) \tag{3.8}$$

3.3 Considerações Finais

Neste capítulo, apresentamos o método utilizado para extrair informações semânticas de um modelo *word embedding* pré treinado para aprimorar a representação de documentos. Essa estratégia, que denominamos de CluWords, representa cada termo em um documento como um conjunto de termos no vocabulário que apresentam uma similaridade em comum. Para isso, apresentamos o processo de geração das CluWords e uma modificação do TF-IDF para explorar esta representação.

Na geração das CluWords, mostramos como extrair, para cada termo, o conjunto de termos semelhantes de acordo com a similaridade de cosseno dos vetores de termos representados pelo modelo *word embedding* pré treinado. Para isso, precisamos determinar um parâmetro α que permite controlar a inclusão de um termo nesse conjunto. Geramos uma matriz de CluWords, onde, cada termo tem sua CluWord correspondente nessa matriz.

A partir da matriz de CluWords, mostramos uma modificação da técnica TF-IDF para classificar as CluWords a serem inclusas na representação BOW estendida de um documento. As CluWords foram criadas com base na semelhança semântica dos termos de modo que a métrica TF-IDF convencional não é capaz de ponderar esses recursos. Assim, demonstramos como combinar os aspectos da métrica TF-IDF convencional com a informação semântica de uma CluWord.

Com isso, concluímos a metodologia que expande a representação de BOW para incluir informações semânticas de acordo com um modelo *word embedding* pré treinado. Com essa nova representação, podemos explorar uma extensa gama de contextos na área de Processamento de Linguagem Natural (PLN), sendo que, nesse trabalho focamos nas subáreas de PLN de Modelagem de Tópicos e Classificação de Documentos.

4 Avaliação Experimental

Nesse capítulo, apresentamos e discutimos uma extensa avaliação experimental para comprovar a eficácia das CluWords em superar as técnicas de Modelagem de Tópicos existentes atualmente. Começamos mostrando a configuração experimental do trabalho, onde, mostramos as bases de dados utilizadas e os métodos de avaliação para comparar os resultados experimentais. Após descrever as bases de dados e métodos utilizados, partimos para os resultados experimentais, onde, exibimos os resultados comparativos que comprovam a eficácia das CluWords em relação aos algoritmos existentes no estado da arte para Modelagem de Tópicos, tornando as CluWords o novo algoritmo de Modelagem de Tópicos estado da arte de acordo com nossos experimentos.

4.1 Configuração Experimental

Nessa seção descrevemos as configurações da nossa avaliação. São utilizadas 12 bases de dados com diferentes características, sendo essa diferença crucial para comprovarmos a eficácia do nosso método. Em seguida, descrevemos os métodos de avaliação (i.e., métricas de Modelagem de Tópicos) que utilizamos para comparar as CluWords com os outros algoritmos de Modelagem de Tópicos existentes no estado da arte. Ainda nos métodos de avaliação, descrevemos a quantidade de tópicos, e palavras principais por tópicos (i.e., *top words*), definidas para cada base de dados e quais os testes estatísticos que utilizamos para comprovar a significância estatística de nossos resultados.

Bases de Dados

O principal objetivo da nossa solução é executar efetivamente a Modelagem de Tópicos para que tópicos mais coerentes sejam extraídos. Para avaliar a coerência do nosso modelo, consideramos 12 bases de dados conhecidas na área de PLN como referência. Duas dessas bases foram criadas por nós, coletando comentários do Facebook e do Uber Apps na Google Play Store. As outras bases foram obtidas de trabalhos anteriores na literatura. Para todas as bases, realizamos a remoção de palavras irrelevantes (usando a lista SMART padrão) e removemos palavras como advérbios, usando o dicionário VADER (GILBERT, 2014), pois a grande maioria das palavras importantes para identificar tópicos são substantivos e verbos. Esses procedimentos melhoram tanto a eficiência quanto a eficácia de todas as estratégias analisadas. A Tabela 2 fornece as informações de cada base de dados de referência, informando o número de atributos (i.e., palavras), documentos, o número médio de palavras por documento (i.e., densidade) e as referências correspondentes.

Tabela 2 – Características das bases de dados.

Dataset	#Documentos	#Palavras	Densidade
20NewsGroup [a]	15,411	29,842	76.408
ACM (VIEGAS, 2015)	22,384	16,811	30.428
Angrybirds (GUZMAN; MAALEJ, 2014)	1,428	1,903	7.135
Dropbox (GUZMAN; MAALEJ, 2014)	1,909	2,430	9.501
Evernote (GUZMAN; MAALEJ, 2014)	8,273	6,307	11.002
Facebook	12,297	5,168	6.427
InfoVisVast [b]	909	6,104	86.215
Pinterest (GUZMAN; MAALEJ, 2014)	3,168	2,174	4.478
TripAdvisor (GUZMAN; MAALEJ, 2014)	2,816	3,152	8.532
Tweets (LI, 2016)	12,030	8,029	4.450
Uber (ZHANG, 2017)	11,541	5,517	7.868
WhatsApp (GUZMAN; MAALEJ, 2014)	2,956	1,777	3.103

[a] http://qwone.com/~jason/20Newsgroups/
[b] http://www.cc.gatech.edu/gvu/ii/jigsaw/datafiles.html

Avaliação, Algoritmos e Procedimentos

Comparamos as estratégias de Modelagem de Tópicos usando métricas da literatura que avaliam a qualidade representativa dos tópicos (NIKOLENKO, 2016; NIKOLENKO, 2017). Em geral, existem três classes de métricas para qualidade de tópicos baseadas em três critérios: (a) *Coherence*, (b) *Mutual Information* e (c) Representação Semântica. Neste trabalho, nos concentramos em (a) e (b), uma vez que são as métricas mais usadas na literatura (SHI, 2018; NIKOLENKO, 2017). Também consideramos três comprimentos dos tópicos (5, 10 e 20 palavras) sob cada métrica em nossa avaliação – diferentes comprimentos podem trazer diferentes desafios.

Coherence captura a facilidade de interpretação de acordo com a coocorrência das palavras. Palavras que coocorrem frequentemente em contextos similares em um corpus são mais fáceis de se correlacionar, pois, geralmente, definem um "conceito" ou "tópico" mais bem definido. Para a métrica *Coherence*, empregamos uma versão melhorada que é usada por Nikolenko, Koltcov e Koltsova (2017). Esta versão, denominada *TF-IDF Coherence*, é definida na Equação 4.1.

$$c_{\text{tf-idf}}(t, W_t) = \sum_{w_1, w_2 \in W_t} log \frac{\sum_{d: w_1, w_2 \in d} \text{tf-idf}(w_1, d) \times \text{tf-idf}(w_2, d)}{\sum_{d: w_1 \in d} \text{tf-idf}(w_1, d)} \tag{4.1}$$

onde a métrica *tf-idf* é calculada com frequência aumentada de acordo com a Equação 4.2.

$$\text{tf-idf}(w, d) = \left(\frac{1}{2} + \frac{f(w, d)}{max_{w' \in d} f(w', d)}\right) \times log\left(\frac{|D|}{|\{d \in D : w \in d\}|}\right) \tag{4.2}$$

e $f(w, d)$ é o número de ocorrências de um termo w no documento d. Isso faz com que

a métrica favoreça os tópicos com altos valores de *tf-idf*, pois o numerador da fração de *Coherence* tem dependência quadrática dos valores de *TF-IDF* e o denominador é apenas linear.

Outra classe de métricas de qualidade de tópico é baseada na noção de *Mutual Information* entre as principais palavras de um tópico. A métrica mais conhecida, e utilizada, dessa classe de métricas é a *Pairwise pointwise Mutual Information* (PMI). A métrica PMI mede quanto uma palavra "ganha" de informação dada à ocorrência de outra palavra, levando em consideração as dependências entre as palavras. Para esta classe de métricas, utilizamos a versão normalizada, denominada de Normalized PMI (NPMI), apontada por 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 4.3.

$$\text{NPMI}_t = \sum_{i<j} \frac{log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-log\ p(w_i, w_j)} \tag{4.3}$$

Em seguida, comparamos nossa representação de dados que foi descrita na Seção 3, bem como a melhor configuração com oito estratégias de Modelagem de Tópicos propostas recentemente, marcadas em negrito na Seção 2. Em nossos experimentos, adotamos a técnica *Non-negative Matrix Factorization* (NMF) (LEE; SEUNG, 1999) para avaliar a CluWords, uma vez que é a principal técnica de fatoração de matriz não probabilística. Descobrimos 25 tópicos para todas as bases de dados, exceto 20News, ACM e Tweets, nos quais foram encontrados tópicos de 20, 11 e 6 para essas bases de dados, respectivamente. O número de tópicos para as bases de dados de aplicativos foi definido com base nas escolhas feitas em (GUZMAN; MAALEJ, 2014). Para as bases 20News,ACM e Tweets, escolhemos o número de tópicos igual ao número de classes de cada base. Avaliamos a significância estatística de nossos resultados por meio do *t-test* pareado (RUXTON, 2006) com 95% de confiança e a correção de Holm-Bonferroni (RICE, 1989) para contabilizar múltiplos testes. Na próxima seção, apresentamos os resultados dos experimentos conduzidos para avaliar a efetividade das CluWords usando três diferentes espaços *word embeddings* pré treinados, considerando os valores de NPMI. Em seguida, comparamos a melhor instanciação de *word embeddings* com as estratégias de linha de base, com relação aos valores de *Coherence* e NPMI.

4.2 Resultados Experimentais

Nessa seção exibimos os resultados que comprovam a eficácia das CluWords na tarefa de Modelagem de Tópicos. Primeiro, definimos qual é o melhor espaço de *word embedding* (Word2Vec – Google News, FastText – WikiNews ou FastText – Common

Crawl) para gerar as CluWords em cada base de dados. Para definir qual o melhor espaço, apresentamos resultados quantitativos de similaridade de cosseno entre as palavras e os resultados de NPMI para cada base de dados. Com isso, comprovamos que o FastText – WikiNews sempre alcança melhores resultados superiores considerando o nosso contexto. A partir do espaço escolhido, realizamos a avaliação da eficácia das CluWords em comparação com as linhas de base. Com isso, conseguimos, a partir das métricas escolhidas e do teste estatístico, comprovar que as CluWords sempre superam as técnicas existentes atualmente na área de Modelagem de Tópicos. Por fim, mostramos como a nossa estratégia pode melhorar a classificação de documentos, comprovando que a as CluWords superam os outros algoritmos em tarefas de classificação automática.

Escolhendo o Melhor Espaço Word Embedding

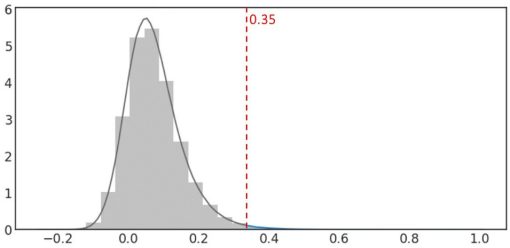
Nesta parte, comparamos a técnica proposta com três *word embeddings* pré treinados: (i) Word2Vec – modelo traindo com o GoogleNews (MIKOLOV, 2013); (ii) FastText – modelo treinado com WikiNews (MIKOLOV, 2017) e (iii) FastText – modelo treinado em *Common Crawl* (MIKOLOV, 2017). Inicialmente, para construir a representação de dados proposta, conforme descrito no Capítulo 3, 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 pré treinados *word embeddings*. Assim, para nossos experimentos, escolhemos um limite α capaz de selecionar apenas 2% dos pares de palavras similares. Selecionamos um limite alto de α apenas para evitar um par inesperado de termos, pois o espaço dos vetores de palavras não é uniformemente distribuído, de acordo com (MIMNO; THOMPSON, 2017). Assim, 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.

A Figura 2 mostra os resultados das CluWords nos três espaços *word embedding* avaliados. O FastText WikiNews sempre alcança resultados superiores considerando todas as bases de dados e tamanhos de tópicos (5, 10 e 20 palavras). 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 Modelagem de Tópicos 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.

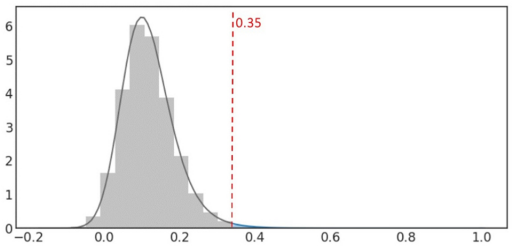
Realizamos um experimento quantitativo adicional usando o espaço WikiNews do FastText para reforçar a evidência de que as CluWords também podem capturar informa-

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

(a) Word2Vec Google News.



(b) FastText WikiNews



(c) FastText Common Crawl

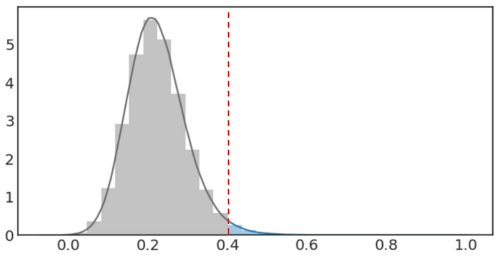


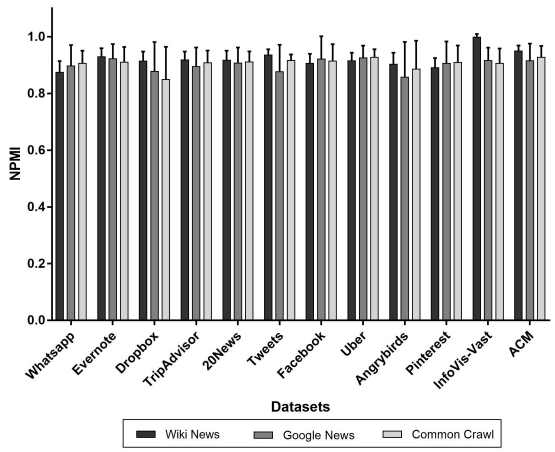
Tabela 3 – Informações sintáticas nas CluWords.

Bases de Dados	**Informação Sintática**
20News	14.47 ± 6.54
Angrybirds	10.76 ± 6.11
ACM	15.58 ± 7.54
Dropbox	12.83 ± 6.53
Evernote	14.02 ± 7.05
Facebook	12.62 ± 6.78
InfoVis-Vast	17.57 ± 7.66
Pinterest	12.12 ± 5.99
Tripadvisor	12.92 ± 6.90
Tweets	13.34 ± 6.37
Uber	12.99 ± 7.05
Whatsapp	10.37 ± 6.21

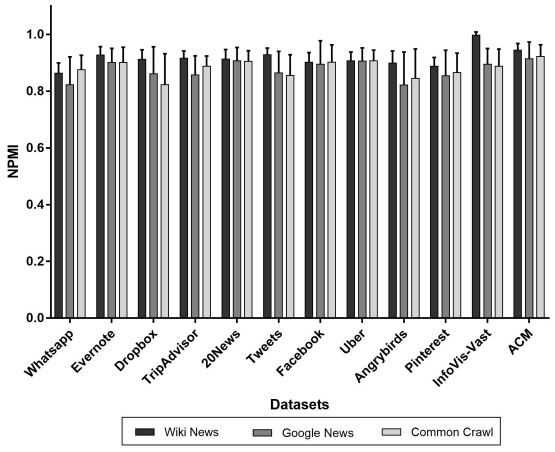
ções sintáticas. No experimento, nosso objetivo é mostrar que no processo de selecionar a vizinhança de uma CluWord C_w (Capítulo 3), uma parte dos termos mais próximos do termo w são variações de uma mesma palavra (e.g., a palavra *chats* é uma variação da palavra *chat*). Assim, dado uma CluWord C_w, selecionamos cada termo $w'|C_{w,w'} \neq 0$ e derivamos w' para seu radical (i.e., *stem*). Medimos a proporção de termos afetados pelo processo de *stemming*. A Tabela 3 ilustra a média dos termos afetados nas CluWords, para as 12 bases de dados. Podemos observar que, aproximadamente, 11% dos termos pertencentes a uma CluWord são uma variação da mesma palavra.

Figura 2 – Avaliação das CluWords explorando diferentes *word embeddings*, em termos da métrica NPMI.

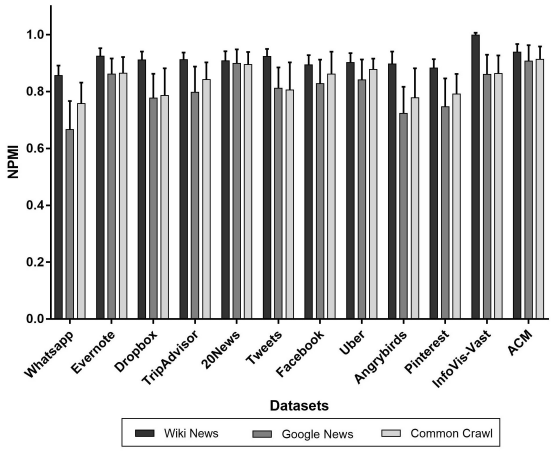
(a) Avaliação com 5 palavras.



(b) Avaliação com 10 palavras.



(c) Avaliação com 20 palavras.



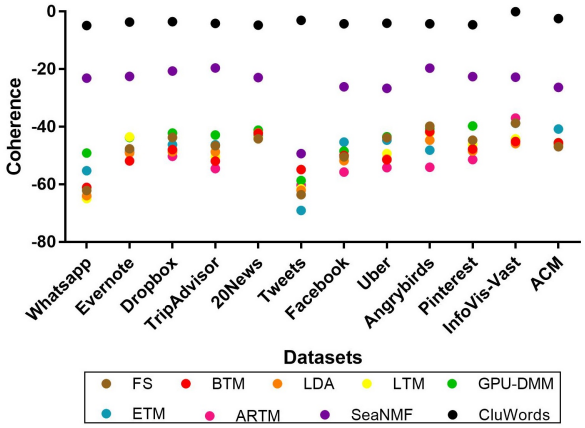
Resultados de Eficácia em Comparação com as Linhas de Base

Nós comparamos nossa solução proposta com oito estratégias de Modelagem de Tópicos no estado da arte, considerando as doze base de dados de referência. Na Figura 3, nossa estratégia obtém ganhos estatisticamente significativos em termos da qualidade dos tópicos descobertos nas 12 bases 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 em mais de 33% dos casos, considerando os três comprimentos de tópicos avaliados. Estes são resultados muito fortes, pois o SeaNMF é considerado o estado da arte em Modelagem de Tópicos (além de ser uma proposta muito recente (SHI, 2018)).

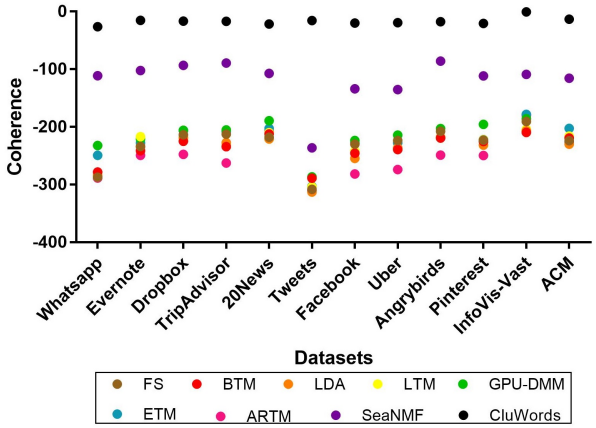
Na Tabela 4, mostramos os resultados das CluWords e as estratégias de comparação escolhidas, considerando a métrica NPMI. Os melhores resultados, marcados com ▲, são estatisticamente superiores aos outros. Os empates estatísticos são representados por

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

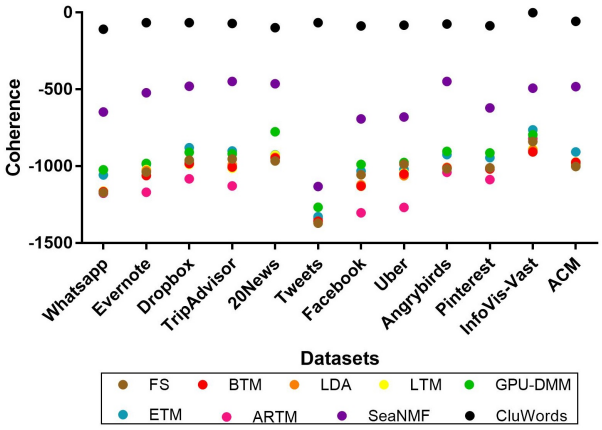
(a) Resultados com 5 palavras.



(b) Resultados com 10 palavras.



(c) Resultados com 20 palavras.



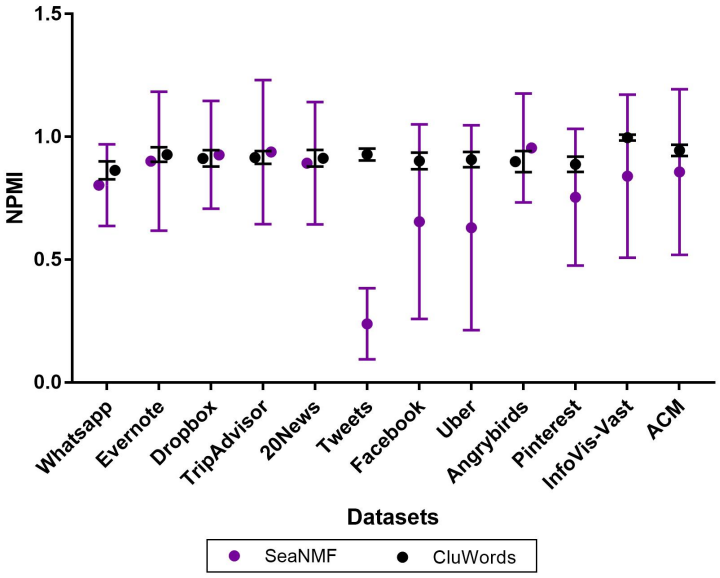
●. Como podemos ver, nossa estratégia atinge os melhores resultados em 7 dos 36 resultados, empatando com o SeaNMF nos outros 29 casos como o método **melhor** em termos de qualidade dos tópicos descobertos, considerando a pontuação NPMI. Novamente, os resultados das outras linhas de base estão muito inferiores, reforçando que o SeaNMF era a linha de base a ser vencida.

Outra perspectiva dos resultados pode ser obtida ao analisar os desvios padrão dos resultados. Os obtidos pelas CluWords são consideravelmente menores que os do SeaNMF. A Figura 4 nos permite observar mais claramente as diferenças entre os desvios para a métrica NPMI, considerando tópicos com 10 palavras. A partir desta figura, podemos observar que, para todas bases de dados, as CluWords obtém valores da métrica NPMI e desvios padrão consideravelmente menores que o SeaNMF. Essa análise comprova a eficácia da estratégia que propomos em bases de dados com diferentes características e, ainda, comprova que as CluWords apresentam melhores resultados que o SeaNMF para as bases de dados avaliadas.

Tabela 4 – Comparação dos resultados obtidos por cada estratégia, considerando as 5, 10 e 20 palavras principais para o NPMI.

Estratégia	**Whatsapp**			**Evernote**		
	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
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
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
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
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
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
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
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 ●
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 ●
Estratégia	**Dropbox**			**TripAdvisor**		
	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras
FS	0.109 ± 0.042	0.097 ± 0.027	0.107 ± 0.018	0.094 ± 0.037	0.092 ± 0.028	0.104 ± 0.021
BTM	0.155 ± 0.050	0.161 ± 0.043	0.166 ± 0.040	0.130 ± 0.052	0.144 ± 0.044	0.158 ± 0.039
LDA	0.165 ± 0.110	0.149 ± 0.056	0.150 ± 0.037	0.114 ± 0.057	0.122 ± 0.029	0.137 ± 0.028
LTM	0.167 ± 0.072	0.160 ± 0.040	0.175 ± 0.046	0.149 ± 0.059	0.144 ± 0.035	0.161 ± 0.037
GPU-DMM	0.284 ± 0.147	0.267 ± 0.125	0.284 ± 0.129	0.286 ± 0.209	0.253 ± 0.144	0.244 ± 0.122
ETM	0.403 ± 0.094	0.399 ± 0.109	0.398 ± 0.119	0.347 ± 0.151	0.349 ± 0.154	0.355 ± 0.163
ARTM	0.158 ± 0.041	0.183 ± 0.036	0.239 ± 0.030	0.128 ± 0.042	0.168 ± 0.030	0.226 ± 0.030
SeaNMF	0.968 ± 0.222 ●	0.927 ± 0.219 ●	0.784 ± 0.185 ●	0.951 ± 0.292 ●	0.938 ± 0.293 ●	0.816 ± 0.262 ●
CluWords	0.914 ± 0.034 ●	0.912 ± 0.033 ●	0.912 ± 0.029 ●	0.918 ± 0.030 ●	0.916 ± 0.026 ●	0.912 ± 0.025 ●
Estratégia	**20News**			**Tweets**		
	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras
FS	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	0.244 ± 0.117	0.217 ± 0.089	0.192 ± 0.059	0.142 ± 0.061	0.100 ± 0.026	0.095 ± 0.019
LDA	0.218 ± 0.121	0.196 ± 0.084	0.174 ± 0.063	0.083 ± 0.055	0.060 ± 0.028	0.079 ± 0.020
LTM	0.224 ± 0.134	0.196 ± 0.074	0.179 ± 0.049	0.109 ± 0.060	0.084 ± 0.022	0.093 ± 0.017
GPU-DMM	0.421 ± 0.044	0.477 ± 0.044	0.471 ± 0.031	0.090 ± 0.062	0.081 ± 0.051	0.092 ± 0.046
ETM	0.249 ± 0.109	0.262 ± 0.092	0.243 ± 0.066	0.057 ± 0.044	0.071 ± 0.038	0.092 ± 0.041
ARTM	0.281 ± 0.105	0.235 ± 0.076	0.216 ± 0.062	0.091 ± 0.055	0.068 ± 0.031	0.080 ± 0.025
SeaNMF	0.897 ± 0.247 ●	0.893 ± 0.249 ●	0.891 ± 0.254 ●	0.237 ± 0.183	0.239 ± 0.145	0.195 ± 0.056
CluWords	0.917 ± 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**		
	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
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
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
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
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
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
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
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
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 ▲
Estratégia	**Angrybirds**			**Pinterest**		
	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras
FS	0.053 ± 0.036	0.077 ± 0.033	0.124 ± 0.028	0.102 ± 0.077	0.096 ± 0.050	0.112 ± 0.031
BTM	0.132 ± 0.075	0.154 ± 0.034	0.193 ± 0.040	0.148 ± 0.074	0.144 ± 0.043	0.147 ± 0.032
LDA	0.137 ± 0.065	0.154 ± 0.038	0.190 ± 0.044	0.144 ± 0.062	0.135 ± 0.043	0.147 ± 0.039
LTM	0.117 ± 0.061	0.154 ± 0.041	0.189 ± 0.041	0.145 ± 0.061	0.137 ± 0.051	0.143 ± 0.035
GPU-DMM	0.260 ± 0.173	0.286 ± 0.141	0.301 ± 0.142	0.330 ± 0.192	0.322 ± 0.194	0.278 ± 0.146
ETM	0.366 ± 0.089	0.373 ± 0.079	0.385 ± 0.085	0.358 ± 0.114	0.355 ± 0.109	0.370 ± 0.115
ARTM	0.209 ± 0.066	0.262 ± 0.054	0.337 ± 0.051	0.167 ± 0.060	0.198 ± 0.030	0.264 ± 0.027
SeaNMF	0.964 ± 0.238 ●	0.955 ± 0.222 ●	0.808 ± 0.194 ●	0.836 ± 0.311 ●	0.754 ± 0.278 ●	0.552 ± 0.167
CluWords	0.903 ± 0.041 ●	0.899 ± 0.043 ●	0.897 ± 0.044 ●	0.891 ± 0.034 ●	0.888 ± 0.031 ●	0.883 ± 0.031 ▲
Estratégia	**InfoVis-Vast**			**ACM**		
	5 palavras	10 palavras	20 palavras	5 palavras	10 palavras	20 palavras
FS	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	0.193 ± 0.079	0.170 ± 0.071	0.149 ± 0.051	0.176 ± 0.084	0.146 ± 0.055	0.136 ± 0.051
LDA	0.154 ± 0.075	0.153 ± 0.064	0.139 ± 0.051	0.138 ± 0.062	0.122 ± 0.051	0.117 ± 0.046
LTM	0.182 ± 0.092	0.158 ± 0.058	0.131 ± 0.042	0.173 ± 0.095	0.163 ± 0.074	0.143 ± 0.054
GPU-DMM	0.264 ± 0.155	0.259 ± 0.104	0.207 ± 0.103	0.233 ± 0.101	0.208 ± 0.113	0.189 ± 0.095
ETM	0.304 ± 0.163	0.304 ± 0.157	0.313 ± 0.158	0.266 ± 0.086	0.230 ± 0.052	0.201 ± 0.035
ARTM	0.102 ± 0.095	0.084 ± 0.077	0.076 ± 0.045	0.178 ± 0.088	0.147 ± 0.058	0.149 ± 0.047
SeaNMF	0.861 ± 0.321 ●	0.840 ± 0.332 ●	0.768 ± 0.288	0.843 ± 0.336 ●	0.857 ± 0.337 ●	0.860 ± 0.345 ●
CluWords	0.998 ± 0.012 ●	0.997 ± 0.012 ●	0.998 ± 0.009 ▲	0.950 ± 0.019 ●	0.945 ± 0.023 ●	0.939 ± 0.028 ●

Figura 4 – Comparação dos valores da métrica NPMI para as estratégias CluWords e SeaNMF considerando 10 palavras



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 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 (LEVENE, 1961) e o teste de Bartlett (BARTLETT, 1937) podem ser usados para verificar essa suposição. O teste de Levene é menos sensível 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 (e.g., p-value $< 0,05$) as diferenças obtidas nas variâncias das amostras provavelmente não ocorreram com base na amostragem aleatória de uma população com variâncias iguais, ou seja, variâncias diferentes.

A Tabela 5 apresenta o teste para igualdade de variâncias com relação aos valores da métrica NPMI para as CluWords e o SeaNMF. Marcamos em ▲ os p-values que apresentam diferenças estatisticamente significativas entre as variâncias e usamos ● para quando as duas estratégias têm a mesma variação. Esta tabela mostra que, em 21 dos 24 testes, as CluWods e o SeaNMF possuem variâncias diferentes. A base de dados Tweets é única em que o teste mostrou variações equivalentes entre as estratégias. No entanto, nesta base de dados, as CluWords superam o SeaNMF considerando os três tamanhos de tópicos mostrados na Tabela 4. Assim, podemos concluir que nossa estratégia é capaz de gerar os melhores tópicos semanticamente coesivos, em termos das métricas *TF-IDF Coherence* e NPMI de acordo com os testes de Levene e Barlett.

Tabela 5 – Teste de igualdade de variâncias considerando 20 palavras.

Base de Dados	**Variância**		**p-value**	
	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▲

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 *clustering*. Para este trabalho, 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 para os tópicos avaliarem o impacto do uso de informações exploradas por estratégias de Modelagem de Tópicos na classificação de documentos. Comparamos três tipos de informações extraídas do modelo de tópico:

1. Tópicos gerados pela estratégia CluWords, que são as informações de vetores latentes extraídas da Modelagem de Tópicos;
2. Os tópicos latentes gerados pelo SeaNMF;
3. A representação dos documentos pela técnica CluWord, descrição no Capítulo 3.

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 (KOHAVI, 1995), sendo que, para a classificação foi utilizado o SVM, que é um método de alta qualidade na classificação de textos (FAN, 2008). 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 (HOLM, 1979) para contabilizar vários testes. Este teste

Tabela 6 – 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

assegura que os melhores resultados, marcados com ▲, são estatisticamente superiores aos outros.

A Tabela 6 apresenta a eficácia da classificação utilizando as métricas Micro-F1 e Macro-F1 (YANG, 1999). Em todas situações, o uso dos tópicos latentes da técnica CluWord obteve os melhores resultados de classificação, com significância estatística nas duas bases de dados avaliadas. Observando a tabela, é possível ver que somente as CluWords individualmente já superam as técnicas BOW e a estratégia utilizando SeaNMF mais tópicos latentes (Tópicos SeaNMF). Além disso, a representação utilizando CluWord mais tópicos latentes (Tópicos CluWords) pode evitar o custo de incluir as CluWords individualmente para representar os documentos. Em outras palavras, ao invés de utilizar as CluWords individualmente – que possui recursos de alta dimensão – é possível utilizar os tópicos latentes gerados a partir das CluWords para a tarefa de classificação automática de documentos. Utilizando os Tópicos CluWords, conseguimos os melhores resultados na tarefa de classificação, onde, quando comparado com o BOW original, foram observados até 10% de ganhos nas métricas para a base de dados ACM. Essa análise indica que as informações fornecidas pela técnica CluWord podem melhorar os resultados da classificação automática de documentos, seja individualmente ou com a estratégia de Modelagem de Tópicos.

4.3 Considerações Finais

Neste capítulo, realizamos uma extensa avaliação experimental para comprovar a eficácia da nossa estratégia em diferentes contextos. Para isso, escolhemos as bases de dados mais utilizadas na literatura na área de Modelagem de Tópicos e os métodos de avaliação para comparar os diferentes algoritmos com as CluWords, sendo essas escolhas baseadas no trabalho recentemente publicado por Viegas, Luiz, Gomes, Khatibi, Canuto, Mourão, Salles, Rocha e Gonçalves (2018).

Após definir a configuração experimental, escolhemos o melhor espaço *word embedding* para criar as CluWords para as bases de dados escolhidas. Para este fim, comparamos

as questões de similaridades que cada espaço podia cobrir e os resultados da métrica NPMI para cada espaço em todas bases de dados. Em todos espaços de *word embeddings* obtivemos resultados satisfatórios, sendo o espaço FastText – WikiNews o que obteve os melhores resultados. Assim sendo, optamos pela escolha do espaço FastText – WikiNews para realizar os experimentos.

Com o espaço *word embedding* escolhido, criamos as CluWords para cada base de dados e avaliamos a Modelagem de Tópicos em comparação com as linhas de bases escolhidas. Comparamos as CluWords com 8 linhas de base, onde, não só superamos a maioria, como também conseguimos melhores o caso de classificação automática de documentos. Com os experimentos realizados, conseguimos cobrir uma grande gama de avaliações e, com isso, conseguimos comprovar que a técnica CluWord é o novo estado da arte, sendo o novo método a ser superado na literatura de PLN.

5 Conclusão e Trabalhos Futuros

Finalizando o trabalho, sumarizamos a estratégia criada e os resultados obtidos por ela, sendo que, nossa estratégia cumpriu todos os objetivos definidos para este trabalho. Por fim, apresentamos os trabalhos futuros, mostrando como as CluWords podem ser utilizadas para tentar aprimorar técnicas existentes na área de Processamento de Linguagem Natural.

5.1 Conclusão

Nesse trabalho, apresentamos uma nova representação de documentos para Modelagem de Tópicos – CluWords. Nossa solução pode ser pensada como o "melhor de todos mundos" (i.e., *best of all worlds*): (i) como dicionários semânticos construídos manualmente, explora relações semânticas explícitas entre palavras, mas sem suas limitações (escalabilidade, adaptabilidade); (ii) explora grandes espaços de *word embedding* para automatizar o processo de computação desses relacionamentos; (iii) conjuga em uma única representação informações sintáticas e semânticas; e (iv) como a estratégia TF-IDF amplamente utilizado, ele propõem uma maneira de medir (i.e., ponderar) a importância de uma dada CluWord para expressar os tópicos de um documento.

Nossa minuciosa avaliação experimental (12 bases de dados, 8 linhas de base, 2 métricas de avaliação, 3 comprimentos de tópicos) mostrou que na maioria das vezes, com grandes margens de diferença, superamos os melhores (i.e., *state-of-art*) métodos para Modelagem de Tópicos conhecidos na literatura, com uma variabilidade muito menor em termos de qualidade dos tópicos produzidos. Em outras palavras, nossos resultados das CluWords os atuais a serem vencidos, estabelecendo um novo padrão elevado para a área de pesquisa Modelagem de Tópicos. Também demonstramos que os tópicos gerados têm o potencial de melhorar outras aplicações, como a classificação automática de texto.

5.2 Trabalhos Futuros

Imaginamos muitos trabalhos futuros à nossa frente. Muito ainda precisa ser entendido em termos das propriedades teóricas dos *clusters* de uma CluWord. Por exemplo, podemos explorar outras noções de similaridade que não dependam apenas dos espaços de incorporação gerados ou usar outros tipos de similaridade mais adequados a esses espaços, evitando resultados inesperados (MIMNO; THOMPSON, 2017)? Como observamos, também há espaço para filtrar tópicos irrelevantes ou retirar palavras não relacionadas das CluWords. Também precisamos avaliar algum tipo de medida de *recall* das CluWords,

por exemplo: elas contêm todas as variações sintáticas que deveriam para um determinado centroide? E, em termos de semântica, quando é que vale a pena ou deveríamos "fundir" diferentes CluWords com os centroides iguais ou muito próximos? Vimos que, na prática, as CluWords com o mesmo centroide ocorrem. Esse aspecto de "fusão"é uma variante da estratégia clássica de *K-Means*, mas as CluWords trazem algumas questões particulares e específicas. Por exemplo, no *K-Means*, nenhum *cluster* compartilha o mesmo centroide quando o processo é interrompido. Mas não está claro se essa propriedade deve ser válida para as CluWords, pois a fusão de *clusters* com os mesmos, ou próximos, centroides podem trazer mais ruído ao processo. E, como estamos falando de grupos de palavras, devemos testar quais estratégias, além do clássico *K-Means*, são mais adequadas para elas, por exemplo, *clustering* hierárquico ou baseado em densidade. Finalmente, outra área que pretendemos explorar para aprender, a partir dos novos dados, não apenas novas estratégias de ponderação para as CluWords, mas também novas medidas de similaridade adaptadas para as particularidades de uma base de dados (densidade,número de atributos, etc.).

Referências

AUDEN, W. H. *The Complete Works of WH Auden: Prose. 1939-1948*. [S.l.]: Princeton University Press, 1996. Citado na página 15.

BICALHO, P. V.; CUNHA, T. de O.; MOURAO, F. H. J.; PAPPA, G. L.; MEIRA, W. Generating cohesive semantic topics from latent factors. In: IEEE. *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*. [S.l.], 2014. p. 271–276. Citado 3 vezes nas páginas 15, 26 e 28.

ALGHAMDI, R.; ALFALQI, K. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, Citeseer, v. 6, n. 1, 2015. Citado na página 15.

BLEI, D. M. Probabilistic topic models. *Communications of the ACM*, ACM, v. 55, n. 4, p. 77–84, 2012. Citado na página 15.

ZHAO, W. X.; JIANG, J.; WENG, J.; HE, J.; LIM, E.-P.; YAN, H.; LI, X. Comparing twitter and traditional media using topic models. In: SPRINGER. *European conference on information retrieval*. [S.l.], 2011. p. 338–349. Citado na página 15.

PEDROSA, G.; PITA, M.; BICALHO, P.; LACERDA, A.; PAPPA, G. L. Topic modeling for short texts with co-occurrence frequency-based expansion. In: IEEE. *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.], 2016. p. 277–282. Citado na página 15.

JIN, O.; LIU, N. N.; ZHAO, K.; YU, Y.; YANG, Q. Transferring topical knowledge from auxiliary long texts for short text clustering. In: ACM. *Proceedings of the 20th ACM international conference on Information and knowledge management*. [S.l.], 2011. p. 775–784. Citado na página 15.

XIE, P.; XING, E. P. Integrating document clustering and topic modeling. *CoRR*, abs/1309.6874, 2013. Disponível em: <http://arxiv.org/abs/1309.6874>. Citado na página 15.

RUBIN, T. N.; CHAMBERS, A.; SMYTH, P.; STEYVERS, M. Statistical topic models for multi-label document classification. *Machine learning*, Springer, v. 88, n. 1-2, p. 157–208, 2012. Citado na página 15.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information processing & management*, Elsevier, v. 24, n. 5, p. 513–523, 1988. Citado 4 vezes nas páginas 16, 19, 27 e 31.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado na página 16.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. Indexing by latent semantic analysis. *Journal of the American society for information science*, Wiley Online Library, v. 41, n. 6, p. 391–407, 1990. Citado 3 vezes nas páginas 16, 26 e 28.

PEARSON, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Taylor & Francis, v. 2, n. 11, p. 559–572, 1901. Citado na página 16.

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, Nature Publishing Group, v. 401, n. 6755, p. 788, 1999. Citado 4 vezes nas páginas 16, 25, 27 e 36.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2013. p. 3111–3119. Citado 3 vezes nas páginas 16, 21 e 27.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. Disponível em: <http://arxiv.org/abs/1301.3781>. Citado 5 vezes nas páginas 17, 20, 27, 29 e 37.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543. Citado 3 vezes nas páginas 17, 20 e 21.

MIKOLOV, T.; GRAVE, E.; BOJANOWSKI, P.; PUHRSCH, C.; JOULIN, A. Advances in pre-training distributed word representations. *CoRR*, abs/1712.09405, 2017. Disponível em: <http://arxiv.org/abs/1712.09405>. Citado 6 vezes nas páginas 17, 20, 21, 27, 29 e 37.

TANG, J.; QU, M.; MEI, Q. Pte: Predictive text embedding through large-scale heterogeneous text networks. In: ACM. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.], 2015. p. 1165–1174. Citado na página 17.

ROTHE, S.; EBERT, S.; SCHÜTZE, H. Ultradense word embeddings by orthogonal transformation. *CoRR*, abs/1602.07572, 2016. Disponível em: <http://arxiv.org/abs/1602.07572>. Citado na página 17.

NGUYEN, D. Q.; BILLINGSLEY, R.; DU, L.; JOHNSON, M. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, v. 3, p. 299–313, 2015. Citado na página 17.

LI, C.; DUAN, Y.; WANG, H.; ZHANG, Z.; SUN, A.; MA, Z. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems (TOIS)*, ACM, v. 36, n. 2, p. 11, 2017. Citado 3 vezes nas páginas 17, 25 e 27.

DAS, R.; ZAHEER, M.; DYER, C. Gaussian lda for topic models with word embeddings. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. [S.l.: s.n.], 2015. v. 1, p. 795–804. Citado 4 vezes nas páginas 17, 23, 25 e 27.

CAVNAR, W. B.; TRENKLE, J. M. N-gram-based text categorization. *Ann arbor mi*, Citeseer, v. 48113, n. 2, p. 161–175, 1994. Citado 2 vezes nas páginas 20 e 27.

BARONI, M.; DINU, G.; KRUSZEWSKI, G. 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)*. [S.l.: s.n.], 2014. v. 1, p. 238–247. Citado na página 20.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: *International Conference on Machine Learning*. [S.l.: s.n.], 2014. p. 1188–1196. Citado na página 21.

LEV, G.; KLEIN, B.; WOLF, L. In defense of word embedding for generic text representation. In: SPRINGER. *International Conference on Applications of Natural Language to Information Systems*. [S.l.], 2015. p. 35–50. Citado 2 vezes nas páginas 22 e 27.

SÁNCHEZ, J.; PERRONNIN, F.; MENSINK, T.; VERBEEK, J. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, Springer, v. 105, n. 3, p. 222–245, 2013. Citado 2 vezes nas páginas 22 e 27.

HYVÄRINEN, A.; OJA, E. Independent component analysis: algorithms and applications. *Neural networks*, Elsevier, v. 13, n. 4-5, p. 411–430, 2000. Citado 2 vezes nas páginas 22 e 27.

PERRONNIN, F.; DANCE, C. Fisher kernels on visual vocabularies for image categorization. In: IEEE. *2007 IEEE conference on computer vision and pattern recognition*. [S.l.], 2007. p. 1–8. Citado 2 vezes nas páginas 22 e 27.

CHATFIELD, K.; LEMPITSKY, V. S.; VEDALDI, A.; ZISSERMAN, A. The devil is in the details: an evaluation of recent feature encoding methods. In: *BMVC*. [S.l.: s.n.], 2011. v. 2, n. 4, p. 8. Citado 2 vezes nas páginas 22 e 27.

HOFMANN, T. Probabilistic latent semantic analysis. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. [S.l.], 1999. p. 289–296. Citado 3 vezes nas páginas 23, 24 e 27.

CHENG, X.; YAN, X.; LAN, Y.; GUO, J. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge & Data Engineering*, IEEE, n. 1, p. 1–1, 2014. Citado 2 vezes nas páginas 23 e 27.

CHEN, Z.; LIU, B. Topic modeling using topics from many domains, lifelong learning and big data. In: *International Conference on Machine Learning*. [S.l.: s.n.], 2014. p. 703–711. Citado 2 vezes nas páginas 23 e 27.

VORONTSOV, K.; POTAPENKO, A. Additive regularization of topic models. *Machine Learning*, Springer, v. 101, n. 1-3, p. 303–323, 2015. Citado 2 vezes nas páginas 24 e 27.

QIANG, J.; CHEN, P.; WANG, T.; WU, X. Topic modeling over short texts by incorporating word embeddings. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2017. p. 363–374. Citado 2 vezes nas páginas 24 e 27.

GUZMAN, E.; MAALEJ, W. How do users like this feature? a fine grained sentiment analysis of app reviews. In: IEEE. *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*. [S.l.], 2014. p. 153–162. Citado 4 vezes nas páginas 24, 27, 35 e 36.

GOLUB, G. H.; REINSCH, C. Singular value decomposition and least squares solutions. *Numerische mathematik*, Springer, v. 14, n. 5, p. 403–420, 1970. Citado 2 vezes nas páginas 24 e 27.

NGUYEN, D. Q.; BILLINGSLEY, R.; DU, L.; JOHNSON, M. Improving topic models with latent feature word representations. *CoRR*, v. 3, p. 299–313, 2018. Disponível em: <https://arxiv.org/abs/1810.06306>. Citado 2 vezes nas páginas 25 e 27.

SHI, B.; LAM, W.; JAMEEL, S.; SCHOCKAERT, S.; LAI, K. P. Jointly learning word embeddings and latent topics. In: ACM. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. [S.l.], 2017. p. 375–384. Citado 2 vezes nas páginas 25 e 27.

MAHMOUD, H. *Pólya urn models*. 1. ed. [S.l.]: Chapman and Hall/CRC, 2008. ISBN 1420059831, 9781420059830. Citado na página 25.

SHI, T.; KANG, K.; CHOO, J.; REDDY, C. K. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In: INTERNATIONAL WORLD WIDE WEB CONFERENCES STEERING COMMITTEE. *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. [S.l.], 2018. p. 1105–1114. Citado 4 vezes nas páginas 25, 27, 35 e 39.

CHOO, J.; LEE, C.; REDDY, C. K.; PARK, H. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics*, IEEE, v. 19, n. 12, p. 1992–2001, 2013. Citado na página 26.

GILBERT, C. H. E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*. [s.n.], 2014. Disponível em: <http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf>. Citado na página 34.

VIEGAS, F.; GONÇALVES, M. A.; MARTINS, W.; ROCHA, L. Parallel lazy semi-naive bayes strategies for effective and efficient document classification. In: ACM. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. [S.l.], 2015. p. 1071–1080. Citado na página 35.

LI, Q.; SHAH, S.; LIU, X.; NOURBAKHSH, A.; FANG, R. Tweetsift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding. In: ACM. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. [S.l.], 2016. p. 2429–2432. Citado na página 35.

ZHANG, Y.; GUEGUEN, L.; ZHARKOV, I.; ZHANG, P.; SEIFERT, K.; KADLEC, B. Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In: *SUNw: Scene Understanding Workshop-CVPR*. [S.l.: s.n.], 2017. Citado na página 35.

NIKOLENKO, S. I. Topic quality metrics based on distributed word representations. In: ACM. *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval*. [S.l.], 2016. p. 1029–1032. Citado 2 vezes nas páginas 35 e 36.

NIKOLENKO, S. I.; KOLTCOV, S.; KOLTSOVA, O. Topic modelling for qualitative studies. *Journal of Information Science*, SAGE Publications Sage UK: London, England, v. 43, n. 1, p. 88–102, 2017. Citado na página 35.

RUXTON, G. D. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, Oxford University Press, v. 17, n. 4, p. 688–690, 2006. Disponível em: <http://dx.doi.org/10.1093/beheco/ark016>. Citado na página 36.

RICE, W. R. Analyzing tables of statistical tests. *Evolution*, Wiley Online Library, v. 43, n. 1, p. 223–225, 1989. Citado na página 36.

MIMNO, D.; THOMPSON, L. The strange geometry of skip-gram with negative sampling. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2017. p. 2873–2878. Citado 2 vezes nas páginas 37 e 46.

LEVENE, H. Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, Stanford University Press, p. 279–292, 1961. Citado na página 42.

BARTLETT, M. S. Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A*, The Royal Society, v. 160, n. 901, p. 268–282, 1937. Citado na página 42.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145. Citado na página 43.

FAN, R.-E.; CHANG, K.-W.; HSIEH, C.-J.; WANG, X.-R.; LIN, C.-J. Liblinear: A library for large linear classification. *Journal of machine learning research*, v. 9, n. Aug, p. 1871–1874, 2008. Citado na página 43.

HOLM, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, JSTOR, p. 65–70, 1979. Citado na página 43.

YANG, Y. An evaluation of statistical approaches to text categorization. *Information retrieval*, Springer, v. 1, n. 1-2, p. 69–90, 1999. Citado na página 44.

VIEGAS, F.; LUIZ, W.; GOMES, C.; KHATIBI, A.; CANUTO, S.; MOURÃO, F.; SALLES, T.; ROCHA, L.; GONÇALVES, M. A. Semantically-enhanced topic modeling. In: ACM. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. [S.l.], 2018. p. 893–902. Citado na página 44.