Uso de expressões do domínio na classificação automática de documentos

Ricardo B. Scheicher[1], **Roberta A. Sinoara**[1], **Newton J. Koga**[1], **Solange O. Rezende**[1]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
Avenida Trabalhador São-carlense, 400, 13566-590, São Carlos - SP, Brasil

{ricardoxem, rsinoara}@usp.br, njk@grad.icmc.usp.br, solange@icmc.usp.br

Abstract. *In order to achieve good Text Mining results, the representation model used to structure the text collection must preserve important information hidden in the text documents. In this context, a text representation based on expressions of a specific domain was developed in a previous work. In this paper, we propose a generalization of that representation, allowing the application of the same idea to any domain. The usage of the new representation was evaluated in text classification scenarios, considering documents written in both Portuguese and English. The results suggest that the combination of a bag-of-words and the new representation improves the classification accuracy, especially in cases of more challenging text classification scenarios.*

Resumo. *No processo de Mineração de Textos é importante que a representação adotada para a coleção de documentos preserve as informações importantes contidas nos textos. Nesse contexto, em um trabalho anterior, foi desenvolvida uma representação baseada em expressões de um domínio específico. Para possibilitar o uso dessa ideia em diferentes domínios, neste artigo é proposta uma generalização dessa representação. O uso da nova representação foi avaliado na classificação de documentos, tanto em português quanto em inglês. Os resultados indicam que o uso da combinação de modelos gerados com a bag-of-words e a nova representação melhoram as acurácias obtidas com apenas a bag-of-words, principalmente em cenários de classificação mais complexos.*

1. Introdução

Devido à grande quantidade de documentos textuais sendo produzidos em grande velocidade e armazenados diariamente em meio digital, a análise manual desse conteúdo torna-se inviável. Contudo, existe a necessidade e o interesse em organizar, classificar e, de modo geral, extrair conhecimento desses dados [Bornmann and Mutz 2015]. Nesse sentido, a área de Mineração de Textos tem como objetivo desenvolver técnicas e algoritmos para dar suporte à extração de conhecimento de grandes coleções de documentos [Gupta and Lehal 2009, Aggarwal and Zhai 2012].

Como apresentado na Figura 1, o processo de Mineração de Textos pode ser dividido em cinco etapas: Identificação do Problema, Pré-processamento, Extração de Padrões, Pós-processamento e Utilização do Conhecimento. O foco principal do trabalho apresentado neste artigo está na etapa de Pré-processamento. Uma atividade muito importante nessa etapa é a representação dos documentos textuais. Os documentos devem ser

colocados em um formato estruturado, que servirá de entrada para os algoritmos de extração de padrões. Assim, é importante que o modelo de representação adotado preserve os padrões a serem descobertos na próxima etapa do processo de Mineração de Textos.

Figura 1. Processo de Mineração de Textos [Rezende et al. 2003].

O modelo de representação textual mais tradicional em Mineração de Textos é o modelo espaço vetorial, cuja a coleção de documentos é representada em uma matriz documento-termo [Tan et al. 2005]. Nessa matriz, cada documento é representado em uma linha e as colunas correspondem a termos encontrados na coleção. Na representação mais comum, denominada de *bag-of-words* (BoW), os termos correspondem a palavras individuais encontradas na coleção. Apesar dos excelentes resultados obtidos no uso geral da BoW, a limitação desta técnica está na representação dos termos de forma independente, não considerando as relações entre esses termos, estruturas morfossintáticas, e ainda, aspectos semânticos e significados. A representação dos textos como um conjunto não ordenado de termos, que não considera a sintaxe ou as relações semânticas existentes entre as palavras (como termos sinônimos e hiperônimos), tem como consequência a perda de parte das informações contidas nos textos. Desta forma, a *bag-of-words* pode não ser o modelo mais apropriado para representar documentos em todas as possíveis aplicações da Mineração de Textos.

Além disso, em muitas aplicações, o usuário ou especialista do domínio possui conhecimento sobre o conteúdo da coleção de documentos e tem uma ideia sobre as possíveis classes ou a organização que é esperada como resultado da Mineração de Textos. Considerando esse fato, no trabalho de Marques et al. (2015) foi desenvolvida uma técnica para representação de documentos textuais para dar suporte à extração de conhecimento a partir de artigos sobre desenvolvimento de produtos e serviços no con-

texto da área de Engenharia de Produção. A técnica desenvolvida, denominada de *bag-of-expressions-of-domain* (BoED), é uma representação no modelo espaço vetorial que incorpora aspectos semânticos por meio de informações do domínio, normalmente fornecidas pelo usuário.

Neste artigo, é proposta uma representação generalizada da BoED, chamada de *generalized-bag-of-expressions-of-domain* (gBoED). Essa nova representação permite que a proposta inicial de Marques et al. (2015), que realiza a representação por expressões do domínio de desenvolvimento de produtos e serviços, seja aplicada para representar coleções de documentos em qualquer domínio. Para avaliar a gBoED, foram executados experimentos de classificação de documentos com duas coleções de documentos distintas e considerando três possíveis cenários de classificação para cada coleção. Na avaliação experimental foram utilizados diferentes algoritmos de classificação, considerando tanto a representação BoW quanto a gBoED de forma independente, bem como a combinação de ambas as representações por meio de abordagens de combinação de classificadores.

Este artigo está organizado da seguinte maneira. Na Seção 2 são apresentados os trabalhos relacionados a este, descrevendo o modelo de representação *bag-of-expressions-of-domain* proposta para documentos da área de Engenharia de Produção. Na Seção 3 é apresentada a generalização proposta neste artigo, a representação *generalized-bag-of-expressions-of-domain*. Na Seção 4 é apresentada a avaliação experimental realizada com o objetivo de verificar o impacto da representação gBoED em um processo de Mineração de Textos. Nessa seção são descritas as coleções de textos utilizadas, a configuração dos experimentos e os resultados obtidos. Em seguida, na Seção 5, são apresentadas as conclusões deste trabalho e trabalhos futuros.

2. Trabalhos Relacionados

Dada a importância da etapa de Pré-processamento para a sucesso do processo de Mineração de Textos, diversos trabalhos têm sido desenvolvidos com o objetivo de encontrar modelos de representação que sejam mais adequados às diferentes aplicações do processo. Muitos desses trabalhos utilizam o método *Latent Semantic Indexing* [Aggarwal and Zhai 2012], cujos conceitos principais são expressos em uma mesma dimensão, e assim, visam tratar problemas causados por termos sinônimos ou polissêmicos. Outros trabalhos fazem uso de fontes de conhecimento externo [Xiang et al. 2013, Spanakis et al. 2012, Gabrilovich and Markovitch 2007]. Como exemplo, o método *Explicit Semantic Analysis* [Gabrilovich and Markovitch 2007] utiliza a Wikipédia para representar documentos como um vetor de conceitos. Também são encontrados trabalhos que utilizam informações provenientes de métodos de processamento de língua natural, como entidades nomeadas [Sinoara et al. 2014], classes morfossintáticas [Spanakis et al. 2012, Bekkerman et al. 2007] e papéis semânticos [Sinoara et al. 2016, Ochoa et al. 2013, Shehata et al. 2010, Persson et al. 2009].

Considerando aspectos semânticos e conhecimento prévio que o usuário tem sobre a coleção de documentos, Marques et al. (2015) desenvolveram uma abordagem específica para o domínio de artigos da área de desenvolvimento de produtos, mais especificamente Sistemas Produto-Serviço. Tal trabalho foi realizado com o objetivo de facilitar a análise de artigos da área como parte de um projeto maior, que visa o desenvolvimento e disponibilização de um portal de conhecimentos das áreas de inovação, desenvolvimento

de produtos, gestão do ciclo de vida de produtos e sustentabilidade[1].

Em uma análise exploratória do uso de técnicas de Mineração de Textos na identificação da aplicação prática de métodos e ferramentas de Sistemas Produto-Serviço, Marques et al. (2015) propuseram um novo modelo, chamado de *bag-of-expressions-of-domain* (BoED), para representar a coleção de artigos em questão. Assim como a BoW, a BoED também é uma representação no modelo espaço vetorial. A principal diferença entre as duas representações é que, enquanto na BoW os termos são palavras independentes, na BoED os termos são formados por expressões do domínio. Com isso, os termos carregam informações semânticas obtidas por meio de três listas de termos do domínio, normalmente fornecidas pelo usuário. Tais listas são descritas a seguir.

1. *Lista de métodos e ferramentas:* a primeira lista é composta por nomes de métodos e ferramentas de Sistemas Produto-Serviço e que são de interesse do usuário. Essa lista foi definida como sendo o conjunto M que contém os nomes dos k métodos ou ferramentas e seus respectivos sinônimos (s_i):
   $M = \{m_1(s_{11}, \ldots, s_{1i}), m_2(s_{21}, \ldots, s_{2i}), \ldots, m_k(s_{k1}, \ldots, s_{ki})\}$.
2. *Lista de palavras de aplicação:* a segunda lista é composta por palavras ou expressões que os autores utilizam para indicar que um método ou ferramenta foi aplicado. Essa lista foi definida como sendo o conjunto A que contém as p expressões que indicam a aplicação de um método ou ferramenta e seus respectivos sinônimos (s_i):
   $A = \{a_1(s_{11}, \ldots, s_{1i}), a_{21}(s_{21}, \ldots, s_{2i}), \ldots, a_p(s_{p1}, \ldots, s_{pi})\}$.
3. *Lista de palavras de desenvolvimento teórico:* a terceira lista é composta por palavras ou expressões que os autores utilizam para apresentar o desenvolvimento teórico de um método ou ferramenta em particular. Essa lista foi definida como sendo o conjunto T que contém as q expressões que indicam o desenvolvimento teórico de um método ou ferramenta e seus respectivos sinônimos (s_i):
   $T = \{t_1(s_{11}, \ldots, s_{1i}), t_{21}(s_{21}, \ldots, s_{2i}), \ldots, t_q(s_{q1}, \ldots, s_{qi})\}$.

Segundo Marques et al. (2015), a tarefa de geração das três listas pode ser feita de forma manual ou utilizando técnicas de Mineração de Textos. Quando realizada de forma manual, deve-se selecionar um conjunto de artigos de referência e gerar as três listas. Para a geração por meio de Mineração de Textos, foi explorado o uso de técnicas de reconhecimento de entidades nomeadas e regras de associação. No entanto, apesar dos resultados serem promissores, algumas deficiências foram identificadas na geração automática dessas listas.

Na representação BoED, cada expressão do domínio é composta por um termo da lista M associado a um termo da lista A ou da lista T. As expressões do domínio são buscadas em cada uma das sentenças do documento. A quantidade de ocorrências de cada expressão em cada documento é verificada e a BoED é construída para aquele conjunto de textos. Para ilustrar esse processo e comparar a representação BoED com a tradicional BoW, considere os seguintes documentos hipotéticos (em idioma inglês, seguindo a definição das listas de termos propostas por Marques et al. (2015)):

D1: *This paper proposes a Quality Function Deployment method.*
D2: *This paper proposes a QFD method.*

[1] http://www.portaldeconhecimentos.org.br/

D3: *This paper presents a case study on Quality Function Deployment method.*

Marques et al. (2015) apresenta as seguintes listas de termos geradas manualmente por um especialista do domínio:

1. M = {Analytic Hierarchy Process (AHP), Brainstorming, Computer Aided Design (CAD), Conjoint Analysis, Delphi, Design for Assembly (DFA), Design for Disassembly (DFD), Design Structure Matrix (DSM), Eco-costs/Value Ratio Model (EVR Model), Failure Mode and Effects Analysis (FMEA), Focus Group, Kansei Engineering, Life Cycle Assessment (LCA), Product-Service Blueprint, Quality Function Deployment (QFD), Technology Roadmap (TRM), Theory of Inventive Problem Solving (TRIZ)};
2. A = {use (uses, using, used, usage), apply (applies, applying, applied, application), validate (validates, validating, validated), case study (case research, action research, cases, real case, practical case)};
3. T = {develop (develops, developing, developed), propose (proposes, proposing, proposed), introduce (introduces, introducing, introduced), suggest (suggests, suggesting, suggested), provide (provides, providing, provided)}.

Na Figura 2 são apresentadas as representações BoW e BoED para os três documentos apresentados anteriormente e considerando as três listas apresentadas por Marques et al. (2015). Para a geração da BoW foram removidas algumas *stopwords*.

	paper	proposes	quality	function	deployment	method	QFD	presents	case	study
$D1$	1	1	1	1	1	1	0	0	0	0
$D2$	1	1	0	0	0	1	1	0	0	0
$D3$	1	0	1	1	1	1	0	1	1	1

(a) *bag-of-words* (BoW)

	Quality-Function-Deployment_propose	Quality-Function-Deployment_case-study
$D1$	1	0
$D2$	1	0
$D3$	0	1

(b) *bag-of-expressions-of-domain* (BoED)

Figura 2. Representações dos documentos $D1$, $D2$ e $D3$.

Dadas as matrizes da Figura 2, é possível verificar a diferença entre as representações BoW e BoED. Os documentos $D1$ e $D2$ possuem o mesmo significado, pois QFD é a sigla de *Quality Function Deployment*. No entanto, na representação BoW os dois documentos são representados por vetores bastante distintos. Já na representação BoED, $D1$ e $D2$ são representados pelo mesmo vetor.

A representação BoED foi definida com o objetivo de diferenciar duas classes de documentos da área de Sistemas Produto-Serviço: (i) documentos que apresentam desenvolvimento teórico de métodos ou ferramentas; e (ii) documentos que apresentam aplicações práticas dos mesmos métodos ou ferramentas. Na próxima seção é apresentada uma generalização da representação BoED, tornando-a independente do domínio e independente do número de classes apresentado pela coleção de documentos.

3. Modelo de Representação de Documentos gBoED

Conforme apresentado na seção anterior, a representação BoED foi desenvolvida para representar conjuntos de documentos de uma área específica e com o objetivo específico de distinguir documentos que apresentam um desenvolvimento teórico de documentos que apresentam aplicações práticas. A fim de possibilitar a aplicação dessa ideia de usar expressões do domínio em outros problemas e avaliá-la de maneira mais ampla, é proposta uma generalização da BoED, denominada *generalized-bag-of-expressions-of-domain* (gBoED).

Na representação BoED o foco principal é a geração de expressões do domínio a partir das três listas de termos do domínio apresentadas na Seção 2. Cada expressão é formada por um método ou ferramenta pertencente à lista M, seguido de elemento de uma das duas outras listas (listas A e T). Os elementos da lista A são palavras ou expressões comumente utilizadas quando o autor do documento quer apresentar um uso prático de um método ou ferramenta. Da mesma forma, os elementos da lista T são utilizados quando se tem a apresentação do desenvolvimento teórico de um método ou ferramenta. Portanto, as listas A e T podem ser vistas como listas que possuem termos que são importantes ou relevantes para uma classe de documentos, ou seja, as listas A e T referem-se, respectivamente, à documentos de aplicação e documentos de desenvolvimento teórico.

Assim, para generalizar a construção de uma BoED para qualquer domínio e problema, as listas A e T podem ser generalizadas para um conjunto de listas de identificadores de classe. Dessa forma, para a geração da representação gBoED são consideradas uma lista de termos do domínio e um conjunto de listas de identificadores de classe. Tais elementos são descritos a seguir.

- *Lista de Termos do Domínio (Domain Keywords):* formada por palavras ou expressões que são importantes para aquela coleção de documentos e para a organização ou classificação esperada como resultado do processo de Mineração de Textos.

  $$Domain\_Keywords = \{k_1, k_2, \ldots, k_i\}$$

  Cada elemento da lista $Domain\_Keywords$ é formado por um termo do domínio t e seus sinônimos s, isto é, $k_i = \{t_i\} \cup \{s_1, \ldots, s_j\}$.
- *Conjunto de Listas de Identificadores de Classe (Class Keywords):* formado por uma ou mais listas de palavras ou expressões que estão particularmente ligadas a uma determinada classe e, assim, são consideradas como termos ou palavras-chaves daquela classe. O número de listas de identificadores de classe pode variar de acordo com a coleção de documentos e com o objetivo do processo de Mineração de Textos.

  $$Class\_Keywords\_Set = \{\{ck_{11}, ck_{12}, \ldots, ck_{1j}\}, \ldots, \{ck_{m1}, ck_{m2}, \ldots, ck_{ml}\}\}$$

  Cada elemento da m-ésima lista do conjunto $Class\_Keywords\_Set$, ck_{mj}, é formado por um termo identificador de classe t e seus sinônimos s, isto é, $ck_{ml} = \{t_l\} \cup \{s_1, \ldots, s_p\}$).

Na representação gBoED os atributos, representados pelas colunas da matriz documento-termo, são formados por expressões do domínio criadas a partir da combinação de elementos da lista $Domain\_Keywords$ com elementos de uma das listas do

conjunto $Class\_Keyword\_Set$. Um esquema da representação gBoED para uma coleção de n documentos é apresentado na Figura 3. Vale ressaltar que a gBoED pode ser vista como uma BoED independente de domínio. Sendo assim, para o problema do domínio de Sistemas Produto-Serviço tratado por Marques et al. (2015), as duas representações são equivalentes.

	$k_1\_ck_{11}$	...	$k_1\_ck_{1j}$	...	$k_i\_ck_{11}$	...	$k_i\_ck_{1j}$	...	$k_1\_ck_{m1}$	...	$k_1\_ck_{ml}$	...	$k_i\_ck_{m1}$	...	$k_i\_ck_{ml}$
d_1															
d_2															
⋮															
d_n															

Figura 3. Esquema da representação de coleção de documentos gBoED.

Na próxima seção é apresentada a avaliação experimental realizada com o objetivo de analisar o impacto da gBoED na classificação de documentos.

4. Avaliação Experimental

Os experimentos foram planejados e executados com o objetivo de verificar o impacto da representação gBoED em diferentes cenários de classificação de documentos. Assim, nessa avaliação experimental foram executadas as três etapas centrais do processo de Mineração de Textos. Na etapa de Pré-processamento, as coleções de documentos foram representadas como Bow e gBoED. Na etapa de Extração de Padrões, foram utilizados um conjunto de algoritmos tradicionais na área de Aprendizado de Máquina, presentes na ferramenta Weka [Witten and Frank 2005], além de dois algoritmos de classificação indutiva baseados em redes bipartidas [Rossi et al. 2014, Rossi et al. 2016]. Também foram executados experimentos combinando classificadores gerados pelas duas representações. Na etapa de Pós-processamento, os modelos gerados foram avaliados por meio da medida acurácia.

Nessa seção são apresentadas as coleções de documentos juntamente com as listas de termos consideradas na geração da gBoED, a configuração dos experimentos executados e os resultados obtidos.

4.1. Coleções de Documentos

Os experimentos foram conduzidos utilizando-se duas coleções de textos, sendo que cada coleção possui três versões diferentes geradas por Sinoara et al. (2016). Tais versões podem representar cenários reais de aplicação, nos quais diferentes usuários e situações necessitam de diferentes classificações para uma mesma coleção de textos. Assim, cada versão pode ser vista como um *gold standard* distinto, referente a um objetivo específico de classificação.

A primeira coleção de textos, denominada *Best sports - Top 4* (BS-Top4), é um conjunto de 283 notícias de esportes escritas em português. Cada documento possui a classificação correspondente a um esporte, podendo ser Fórmula 1, Motovelocidade, Futebol ou Tênis. A partir dessa coleção de documentos, foram geradas as versões utilizadas nos experimentos. A primeira versão corresponde à classificação padrão por esporte. A segunda e a terceira versões estão relacionadas ao desempenho dos atletas brasileiros.

Nesse caso, existem quatro possíveis rótulos de classificação: "Brasileiro venceu", "Brasileiro não venceu", "Não foi citado brasileiro" ou "Ruído". Assim, as três versões que representam as possibilidades de classificação para essa coleção BS-Top4 são: (i) *BS-tópico*: classificação por esporte; (ii) *BS-semântico*: classificação por desempenho dos atletas brasileiros; e (iii) *BS-tópico-semântico*: classificação por esporte e desempenho dos atletas.

Para a coleção BS-Top4, o cenário de classificação *BS-semântico* foi considerado para gerar as listas de termos do domínio e de identificadores de classe para a construção da gBoED. Assim, um usuário especialista do domínio forneceu as seguintes listas.

- $Domain\_Keywords$: lista de nomes de atletas brasileiros.
- $Class\_Keywords1$: lista de verbos utilizados para expressar vitórias.
- $Class\_Keywords2$: lista de verbos utilizados para expressar derrotas.

A segunda coleção de textos é denominada SemEval-2015. Ela é composta por textos com opiniões de satisfação de usuários, escritos em inglês, sobre hotéis, restaurantes e *laptops*. Essa coleção foi disponibilizada para a *SemEval-2015 Aspect Based Sentiment Analysis Task* [Pontiki et al. 2015]. Os textos são anotados com a polaridade (positivo, negativo ou neutro) de cada aspecto dos produtos avaliados pelo autor. Para a coleção SemEval-2015, considerou-se três versões distintas: (i) *SE-produto*: classificação pelo tipo de produto (Hotel, Restaurante ou *Laptop*); (ii) *SE-polaridade*: classificação pela polaridade (positiva, negativa ou neutra); e (iii) *SE-produto-polaridade*: classificação pelo tipo de produto e pela polaridade.

Para a coleção SemEval-2015, foram utilizadas as seguintes listas de termos do domínio e de identificadores de classe para a construção da gBoED.

- $Domain\_Keywords$: lista de aspectos de hotéis, restaurantes e *laptops* extraídos dos textos de opinião da coleção SemEval-2015, disponibilizada por Pontiki et al. (2015) .
- $Class\_Keywords1$: lista de palavras positivas para o idioma inglês, originalmente utilizada no trabalho de Hu & Liu (2004).
- $Class\_Keywords2$: lista de palavras negativas para o idioma inglês, originalmente utilizada no trabalho de Hu & Liu (2004)[2].

4.2. Representações Utilizadas

Para a execução desses experimentos, foram geradas as representações BoW e gBoED para as diferentes versões das duas coleções de documentos. Para a geração das representações BoW, foi realizada a remoção de *stopwords*, além da radicalização dos termos.

Para gerar a gBoED, foram utilizadas as listas de termos do domínio e de identificadores de classes apresentadas na seção anterior. As expressões do domínio foram geradas com os termos das listas radicalizados.

4.3. Configuração dos Experimentos

Nessa avaliação experimental, executou-se três conjuntos de experimentos de classificação: (i) classificação indutiva supervisionada utilizando a representação BOW, que foi

[2]As listas de palavras positivas e negativas foram obtidas em: http://www.cs.uic.edu/ liub/FBS/opinion-lexicon-English.rar

utilizada como *baseline*; (ii) classificação indutiva supervisionada utilizando a representação gBoED; e (iii) combinação de classificadores gerados utilizando ambas as representações BoW e gBoED.

Os algoritmos utilizados são apresentados a seguir, seguidos dos parâmetros de configuração utilizados em cada caso. Foram utilizadas as implementações da ferramenta Weka [Witten and Frank 2005] para os primeiros cinco algoritmos.

- ***Naive Bayes* (NB)**;
- ***Multinomial Naive Bayes* (MNB)**;
- **J48**, algoritmo de indução de árvores de decisão. Foi utilizado o valor 0,25 para parâmetro *confidence factor*.
- ***Suport Vector Machine* (SVM)**, algoritmo *Sequential Minimal Optimization* (SMO). Nesse algoritmo foram consideramos três tipos de kernel: linear, polinomial (expoente=2) e RBF (Radial Basis Function). Os valores considerados para cada tipo de kernel foram 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0, 1, 10, 10^2, 10^3, 10^4, 10^5.
- ***K-nearest neighbor* (KNN)**, algoritmo IBk. Foram utilizadas as opções de voto com peso e voto sem peso. Os valores utilizados de k foram 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 25, 35, 45, 55. O algoritmo foi executado com duas opções de medida de distância: distância euclideana e cosseno.
- **IMBHN$^{\mathbf{C}}$** [Rossi et al. 2014] e **IMBHN$^{\mathbf{R}}$** [Rossi et al. 2016], algoritmos de classificação indutiva baseados em redes heterogêneas bipartidas. Nesses algoritmos utilizou-se taxa de correção de erros de 0,01, 0,05, 0,1, 0,5. O número máximo de iterações foi ajustado para 1000 e utilizou-se o erro dos mínimos quadrados com critério de parada de 0,01.

Para realizar a combinação de classificadores gerados com cada representação (BoW e gBoED) foram utilizados três abordagens: (i) uso da resposta do classificador com maior confiança (*Most Confident* (MC)); (ii) uso da resposta com maior soma de confianças (*Sum of Confidences* (SC)); e (iii) uso da resposta com maior soma de confianças ponderadas pela acurácia dos classificadores no conjunto de treinamento (*Weighted Sum of Confidences* (WSC)). Além disso, para as três abordagens também foram utilizados diferentes pesos para as representações. Considere w_1 o peso do classificador gerado com a representação BoW e w_2 o peso do classificador gerado com a gBoED. Nesses experimentos, foram utilizados os seguintes valores para esses pesos: $w_1 = \{0,1; 0,3; 0,5; 0,7; 0,9\}$ e $w_2 = 1 - w_1$.

Os classificadores gerados foram avaliados considerando a medida acurácia (porcentagem de documentos corretamente classificados) obtidas por um processo de validação cruzada (*10-fold cross validation*). Os resultados obtidos são apresentados na próxima subseção.

4.4. Resultados e Discussão

Na Tabela 1 é apresentado o conjunto de melhores acurácias obtidas na execução de cada algoritmo para cada versão das coleções de documentos BS-Top4 e SemEval-2015. São apresentados os resultados obtidos com classificadores gerados a partir das representações BoW e gBoED, bem como com a combinação de classificadores gerados com as duas representações. Para cada algoritmo apresentado na Tabela 1, os valores de acurácia

iguais aos obtidos com a BoW para a mesma versão das coleções de documentos estão sublinhados. Os valores maiores que os obtidos com a BoW estão em negrito e os melhores resultados para cada algoritmo (quando não é o resultado obtido com a BoW) está em negrito e sublinhado.

Tabela 1. Melhores acurácias para as coleções *BS-Top4* e *SE-2015*

			Combinação BoW + gBoED					Combinação BoW + gBoED		
	BoW	**gBoED**	**MC**	**SC**	**SWC**	**BoW**	**gBoED**	**MC**	**SC**	**SWC**
	BS-tópico					*SE-produto*				
IMBHN^C	98.9286	67.2167	98.9286	78.4360	98.9286	98.1587	86.0193	98.1587	**98.5276**	**98.5276**
IMBHN^R	99.6429	68.9532	99.6429	70.6527	99.6429	99.1388	89.6914	99.1388	99.0169	99.0169
J48	96.8227	61.8719	96.8227	58.6823	96.8227	92.2704	80.6158	92.2704	92.2704	92.2704
KNN	99.6552	71.0099	99.6552	99.6552	99.6552	98.4041	88.4688	98.4041	**98.5260**	**98.5276**
MNB	100.0000	78.4360	100.0000	100.0000	100.0000	99.5077	86.2572	99.5077	99.5077	99.5077
NB	99.6429	71.4163	99.6429	99.6429	99.6429	92.7612	84.5348	**94.2307**	**93.8633**	**95.0979**
SMO	100.0000	71.7118	100.0000	100.0000	100.0000	96.4408	89.4550	96.4408	96.3189	96.3189
	BS-semântico					*SE-polaridade*				
IMBHN^C	64.6552	44.2118	64.6552	**65.3818**	**66.4286**	80.4908	65.0316	80.4908	**80.6143**	**80.6128**
IMBHN^R	68.9532	43.1034	68.9532	**69.2857**	68.9409	82.8214	64.2939	82.8214	81.9587	82.0822
J48	58.0049	44.5197	**61.1576**	**61.4655**	**64.6675**	71.5236	66.0132	**72.4977**	**72.4992**	**72.9916**
KNN	65.3818	49.4828	**66.7980**	**69.6059**	**69.2734**	77.4179	68.5953	**93.0051**	**91.0388**	**92.8696**
MNB	59.7414	52.9926	**60.4557**	**63.9778**	**65.7512**	84.5438	66.7525	84.5438	**84.5453**	**84.5453**
NB	57.6108	46.6379	**59.0025**	**59.0271**	**61.1084**	70.3071	63.6886	70.3071	70.1867	**70.3086**
SMO	63.6576	48.8300	63.6576	**64.7167**	64.3596	81.6110	69.0786	81.6110	81.2436	81.3640
	BS-tópico-semântico					*SE-produto-polaridade*				
IMBHN^C	62.5739	42.0443	**62.9310**	**67.0813**	**68.8300**	77.4345	60.1235	**77.5565**	**77.7989**	**78.4056**
IMBHN^R	57.992	38.5099	57.9926	56.5764	56.5764	73.9762	60.7347	73.9762	72.5083	72.9976
J48	54.7906	31.8596	**56.5394**	**55.4926**	**57.2414**	71.0479	55.7106	**71.6667**	**71.6652**	**71.2993**
KNN	65.7512	38.8793	**66.3916**	**66.0591**	**67.1552**	75.9425	60.9861	**76.0659**	75.6986	75.5751
MNB	62.5985	44.1749	**65.0369**	**65.0369**	**64.0025**	83.6811	62.2162	83.6811	**83.8046**	**83.8046**
NB	57.2537	43.8054	57.2537	**57.6108**	**57.6108**	68.9521	57.0671	**69.0696**	**69.5604**	**69.8103**
SMO	66.8596	38.5345	66.8596	**68.9901**	**67.9433**	77.8034	60.8627	77.8034	**78.0473**	**78.0473**

Pode-se notar que a representação gBoED não leva a boas acurácias quando é utilizada de modo independente, sendo que os valores de acurácia obtidos são sempre menores do que as acurácias obtidas com o uso da representação BoW. Porém, quando se observa os modelos gerados com a combinação de classificadores gerados com as duas representações, verifica-se que a combinação melhora os resultados obtidos com a BoW em algumas configurações experimentadas.

Considerando as diferentes versões de cada coleção de textos, verifica-se que as versões *BS-tópico* e *SE-produto* correspondem a cenários de classificação mais fáceis. Estes são os cenários tradicionalmente tratados em problemas da literatura e apenas a representação BoW é suficiente para atingir acurácia próxima a 100%. No entanto, os resultados indicam que a BoW não é suficiente para cenários de classificação mais complexos, representados pelos cenários *BS-semântico*, *BS-tópico-semântico*, *SE-polaridade* e *SE-produto-polaridade*. Nesses cenários, a combinação da BoW com a gBoED levou a melhores valores de acurácia na maioria das configurações testadas. Vale notar a melhora de acurácia obtida com a combinação de classificadores no caso *SE-polaridade* utilizando o algoritmo KNN. Para esse caso, o melhor modelo gerado utilizando a BoW obteve 77,4179% de acurácia, enquanto a combinação de classificadores gerados com as

duas representações obteve 93,0051%.

5. Conclusão

O modelo de representação adotado para coleções de documentos durante o processo de Mineração de Textos tem grande impacto no resultado final do processo. Nesse contexto, uma nova representação de textos é proposta neste artigo. Tal representação, denominada *generalized-bag-of-expressions-of-domain* ou gBoED, é gerada com base em listas de termos do domínio. Com isso, incorpora-se aspectos semânticos e conhecimento do domínio em uma representação no modelo espaço vetorial.

Uma avaliação experimental foi realizada com o objetivo de avaliar o impacto da nova representação na classificação automática de textos. A nova representação se mostrou promissora, principalmente em cenários em que a classificação automática é mais complexa e o uso da BoW apresenta resultados mais baixos.

Os resultados obtidos neste trabalho são coerentes com os resultados obtidos com outra representação semântica, apresentados em Sinoara et al. (2016). Como trabalho futuro, os resultados obtidos serão comparados com outros métodos de representação de documentos, como o uso n-gramas e *Latent Semantic Indexing*. Também espera-se aprimorar a geração das expressões do domínio considerando outros aspectos linguísticos dos textos, como o uso de voz passiva e de negações. Além disso, com o objetivo de facilitar a geração da gBoED para diferentes problemas, pretende-se investigar métodos para geração automática das listas de termos do domínio.

6. Agradecimentos

Os autores agradecem os auxílios fornecidos para o desenvolvimento deste trabalho, processo nº 2013/14757-6, processo nº 2014/08996-0 e processo nº 2016/07620-2, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Referências

Aggarwal, C. C. and Zhai, C., editors (2012). *Mining Text Data*. Springer.

Bekkerman, R., Raghavan, H., Allan, J., and Eguchi, K. (2007). Interactive clustering of text collections according to a user-specified criterion. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 684–689.

Bornmann, L. and Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of International Joint Conference on Artifical Intelligence*, pages 1606–1611.

Gupta, V. and Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1:60–76.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM.

Marques, C. A. N., Matsuno, I. P., Sinoara, R. A., Rezende, S. O., and Rozenfeld, H. (2015). An exploratory study to evaluate the practical application of pss methods and tools based on text mining. In *Proceedings of the 20th International Conference on Engineering Design*, pages 7–311–7–320.

Ochoa, J. L., Valencia-García, R., Perez-Soltero, A., and Barceló-Valenzuela, M. (2013). A semantic role labelling-based framework for learning ontologies from spanish documents. *Expert Systems with Applications*, 40(6):2058–2068.

Persson, J., Johansson, R., and Nugues, P. (2009). Text categorization using predicate-argument structures. In *Proceedings of the Nordic Conference of Computational Linguistics*, pages 142–149.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 486–495.

Rezende, S. O., Pugliesi, J. B., Melanda, E. A., and de Paula, M. F. (2003). Mineração de dados. In Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, pages 307–335. Editora Manole.

Rossi, R. G., Lopes, A. d. A., Faleiros, T. d. P., and S. O. Rezende (2014). Inductive model generation for text classification using a bipartite heterogeneous network. *Journal of Computer Science and Technology*, 29(3):361–375.

Rossi, R. G., Lopes, A. d. A., and Rezende, S. O. (2016). Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. *Information Processing & Management*, 52(2):217–257.

Shehata, S., Karray, F., and Kamel, M. S. (2010). An efficient model for enhancing text categorization using sentence semantics. *Computational Intelligence*, 26(3):215–231.

Sinoara, R. A., Rossi, R. G., and Rezende, S. O. (2016). Semantic role-based representations in text classification. In *Proceedings of the 23rd International Conference on Pattern Recognition. No prelo.*

Sinoara, R. A., Sundermann, C. V., Marcacini, R. M., Domingues, M. A., and Rezende, S. O. (2014). Named entities as privileged information for hierarchical text clustering. In *Proceedings of International Database Engineering & Applications Symposium*, pages 57–66.

Spanakis, G., Siolas, G., and Stafylopatis, A. (2012). Exploiting wikipedia knowledge for conceptual hierarchical clustering of documents. *Computer Journal*, 55(3):299–312.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2nd edition.

Xiang, W., Yan, J., Ruhua, C., and Hua, F. (2013). Improving text categorization with semantic knowledge in wikipedia. *IEICE Transactions on Information and Systems*, 96(12):2786–2794.