Exploração do uso de expressões do domínio na classificação de sentimentos

Ricardo B. Scheicher[1], **Roberta A. Sinoara**[2], **Solange O. Rezende**[1]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)

[2]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Campus Boituva

ricardoxem@usp.br, roberta.sinoara@ifsp.edu.br, solange@icmc.usp.br

Abstract. *Sentiment classification is an text mining application presents great challenges, requiring an adequate treatment of the textual semantic. In this work are presented the results of an experimental evaluation of a sentiment classification method for sentiment classification improvement that uses semantically enriched information. Such representation is built based on domain expressions, providing its interpretability and explainability. The results indicate that the method is promising, increasing accuracy values for databases with more concentrated information on a specific subject or domain.*

Resumo. *A classificação de sentimentos é uma aplicação da mineração de textos que apresenta grandes desafios, necessitando um tratamento adequado da semântica dos textos. Neste trabalho são apresentados os resultados de uma avaliação experimental de um método de classificação de sentimentos que utiliza informações semanticamente enriquecidas. Tal representação é construída com base em expressões de domínio, favorecendo a sua interpretabilidade e explicabilidade. Os resultados indicam que o método é promissor, elevando valores de acurácia para bases de dados cujo domínio é mais específico.*

1. Introdução

A análise de sentimentos é uma das desafiadoras aplicações da mineração de textos e tem o objetivo de organizar e classificar opiniões de usuários (ou clientes) sobre produtos e serviços [Liu 2012]. A classificação de sentimentos é uma das tarefas da análise de sentimentos que visa categorizar textos de acordo com as orientações de sentimentos, sendo que métodos de classificação automática de textos podem ser aplicados nessa tarefa. No entanto, a classificação pela polaridade do sentimento é mais complexa do que as tarefas tradicionais de classificação, como a classificação de tópicos. Na classificação de tópicos, o conjunto de palavras individuais, representado por uma *Bag-of-Words* (BOW), geralmente é suficiente para definir a classe do documento. Por exemplo, em uma coleção de notícias de esportes, a ocorrência frequente de algumas palavras como "pneu", "carro" e "circuito" é suficiente para definir que o documento pertence à categoria de notícias de Fórmula 1. Na classificação de sentimentos, o uso de palavras independentes pode não ser suficiente para classificar corretamente o sentimento expresso em um documento, sendo também necessária a incorporação de informações sobre o significado do texto ou conhecimento do domínio.

Bons resultados de classificação de sentimentos foram obtidos com o uso de *word embeddings* [Ju and Yu 2018, Xiong 2016]. O uso de tais representações com base na

semântica latente atinge bom desempenho em classificação, mas afeta negativamente a interpretabilidade dos recursos de representação de texto e, portanto, a explicabilidade de certos modelos de classificação. Embora o desempenho da classificação seja importante, aspectos de interpretabilidade ou explicabilidade podem ser cruciais para algumas aplicações de mineração de texto [Dosilovic et al. 2018, Goodman and Flaxman 2017]. Assim, existe a necessidade do desenvolvimento de métodos alternativos para a incorporação da semântica na representação e classificação dos textos.

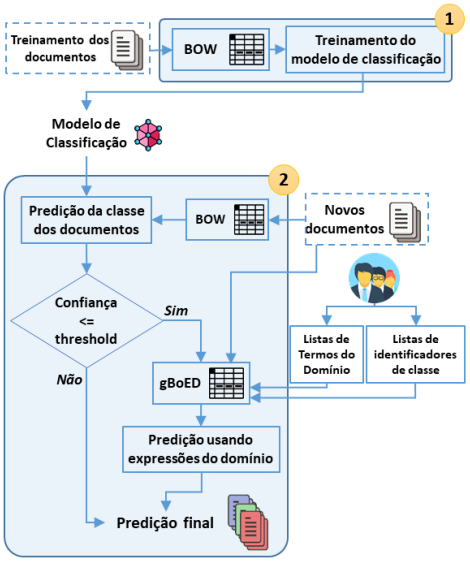
Em [Scheicher et al. 2019] é apresentado um método que visa melhorar resultados de classificação de sentimentos com base em informações semanticamente enriquecidas. Tais informações, denominadas expressões do domínio, agregam o conhecimento do especialista e são concentradas na representação *generalized Bag of Expressions of Domain (gBoED)* [Scheicher et al. 2016]. Uma expressão do domínio é composta pela união de um termo de domínio (TD, termos importantes em um domínio específico) e um identificador de classe (IC, termos importantes para uma classe específica). Por exemplo, em avaliações de restaurantes, "comida" e "serviço" são termos do domínio, "bom" e "terrível" são identificadores das classes positivo e negativo, respectivamente. Portanto, "comida_boa" seria uma expressão de domínio da classe positiva e "serviço_terrível" seria uma expressão de domínio negativa.

Nesse trabalho, é realizada uma extensão do trabalho de [Scheicher et al. 2019], sendo apresentada uma nova avaliação experimental aplicada a diferentes conjuntos de dados. Na Seção 2 será apresentada uma breve explicação sobre o método de classificação e na Seção 3 é apresentada a nova avaliação experimental.

2. Método de Classificação

O método proposto em [Scheicher et al. 2019], apresentado na visa melhorar a classificação de sentimentos utilizando conhecimento do especialista do domínio. Como é possível observar na Figura 1, o método consiste em duas etapas bem definidas. A etapa (1) corresponde ao treinamento de um modelo de classificação baseado na BOW, visto que essa é uma representação simples que alcança bons resultados em diversos cenários de classificação de textos. Na etapa (2), novos documentos são preparados e classificados pelo modelo da etapa (1). Os resultados da classificação cuja confiança foi menor ou igual a um limite global são submetidos a uma nova classificação apoiada pelas informações semanticamente enriquecidas da representação gBoED, caso contrário é mantida a predição do classificador da etapa (1). O limite global está diretamente relacionado à quantidade de documentos que serão selecionados para a reclassificação, influenciando o

Figura 1. Método de melhoria de resultados de classificação. Adaptado de [Scheicher et al. 2019]



resultado final.

O processo de melhoria dos resultados de classificação utiliza a representação gBoED, cujos atributos são expressões do domínio. O especialista é responsável por construir as listas de TD, ID e seus respectivos sinônimos. Também são propostas duas formas de construção da representação gBoED: "*gBoED_Freq*", que considera a frequência das expressões para cada documento; e "*gBoED_Dist*", em que atribi-se um peso de cada expressão do domínio para cada documento, dado pela soma dos inversos multiplicativos do número de palavras entre o TD e o IC em cada sentença.

3. Avaliação Experimental

Nesta seção são apresentados os principais resultados obtidos na avaliação experimental extendida do método de melhoria de resultados de classificação usando expressões do domínio, proposto em [Scheicher et al. 2019]. A descrição dos datasets utilizados, bem como os detalhes da avaliação experimental, incluindo algoritmos e parâmetros utilizados, estão disponibilizados no site[1]. Nos experimentos executados foi utilizado método de amostragem *10-fold cross validation*. Vale notar que a execução dos experimentos realizada nesta nova versão produziu algumas diferenças em relação aos resultados anteriores, devido a atualizações na biblioteca SciKit Learn, da linguagem Python3, disponibilizadas entre a execução desses dois conjuntos de experimentos.

Neste trabalho, foram utilizados os conjuntos de dados HuLiu2004 [Hu and Liu 2004], SemEval2014 [Pontiki et al. 2014] e SemEval2015 [Pontiki et al. 2015], aqueles utilizados em [Scheicher et al. 2019]. Também foi realizada uma segmentação dos conjuntos de dados **SemEval**, separando-as por domínio. A segmentação consiste em **SemEval2014_Laptop** (com 619 avaliações positivas e 622 negativas do domínio de *laptops*), **SemEval2014_Rest** (1.217 avaliações positivas e 451 negativas de restaurantes), **SemEval2015_Hotel** (21 avaliações positivas e 8 negativas de hotéis), **SemEval2015_Laptop** (277 avaliações positivas e 151 negativas de *laptops*), e **SemEval2015_Rest** (257 avaliações positivas e 87 negativas de restaurantes).

Considerando-se os diferentes conjuntos de dados, algoritmos e parâmetros da configuração experimental, foram avaliados um total de 4.575 classificadores. Na Tabela 1 são apresentas as melhores acurácias obtidas para cada conjunto de dados e para cada algoritmo, considerando-se todos os parâmetros testados. São apresentados tanto os resultados obtidos pelo modelo de classificação baseado na BOW quanto pelas duas versões do método proposto (gBoED_Freq e gBoED_Dist). Os valores maiores do que o *baseline* BOW são destacados em negrito, as células em cinza correspondem à melhor acurácia de cada linha e os valores sublinhados indicam o valor mais alto entre gBoED_Freq e gBoED_Dist. O cabeçalho de cada conjunto de dados corresponde aos melhores resultados para cada conjunto de dados.

Analisando os resultados da Tabela 1, observa-se que os resultados de maior destaque foram alcançados para o conjunto de dados *HuLiu2004* e *SemEval2015_Hotel*, melhorando a maior acurácia de classificação da BOW nos 8 algoritmos testados. Em *HuLiu2004*, o melhor valor obtidos foi utilizando gBoED_Freq para o algoritmo SVM-Poly, cujo valor foi de 0,84149. Utilizando gBoED_Dist, o melhor valor resultado foi

[1]Experimentos ERAMIA 2020: http://sites.labic.icmc.usp.br/ricardoxem/eramia2020

Tabela 1. Melhores acurácias para cada conjunto de dados e algoritmo.

	BOW	gBoED_Freq	gBoED_Dist
HuLiu2004	0,82126	**0,84149**	**0,84138**
C4.5-entropia	0,68989	**0,75770**	**0,75759**
C4.5-gini	0,67253	**0,73701**	**0,74379**
KNN-cosseno	0,76057	**0,79057**	**0,79391**
KNN-euclidiano	0,76057	**0,79057**	**0,79391**
MNB	0,81402	**0,83138**	**0,82782**
SVM-linear	0,82126	**0,83828**	**0,84138**
SVM-poly	0,82080	**0,84149**	**0,83851**
SVM-rbf	0,81782	**0,82161**	**0,82115**
SemEval2014	0,80716	0,79856	0,79995
C4.5-entropia	0,63700	**0,63974**	**0,66621**
C4.5-gini	0,63837	**0,64490**	**0,66484**
KNN-cosseno	0,76007	0,75045	0,75904
KNN-euclidiano	0,76557	0,74941	0,75973
MNB	0,80716	0,79720	0,79995
SVM-linear	0,79237	0,79203	0,79203
SVM-poly	0,78824	**0,78859**	**0,78859**
SVM-rbf	0,79925	0,79856	0,79856
SemEval2014_Laptop	0,80017	0,79372	0,79613
C4.5-entropia	0,67125	0,65270	0,65593
C4.5-gini	0,67934	0,65757	0,66079
KNN-cosseno	0,74377	0,73732	**0,74539**
KNN-euclidiano	0,74701	**0,75103**	**0,75909**
MNB	0,80017	0,79372	0,79613
SVM-linear	0,78890	0,78730	**0,78970**
SVM-poly	0,78728	**0,78971**	**0,79535**
SVM-rbf	0,78324	0,78243	**0,78728**
SemEval2014_Rest	0,81416	0,81356	0,81415
C4.5-entropia	0,75003	0,74943	0,74943
C4.5-gini	0,75543	**0,75543**	**0,75603**
KNN-cosseno	0,77282	0,76442	0,77281
KNN-euclidiano	0,77040	0,76141	0,76921
MNB	0,81299	0,80998	0,81178
SVM-linear	0,79018	0,78959	**0,79136**
SVM-poly	0,79377	**0,79377**	**0,79737**
SVM-rbf	0,81416	0,81356	0,81415

	BOW	gBoED_Freq	gBoED_Dist
SemEval2015	0,86520	0,86392	0,86267
C4.5-entropia	0,75903	0,73528	0,73403
C4.5-gini	0,79026	0,78776	0,78776
KNN-cosseno	0,80522	0,78284	0,78281
KNN-euclidiano	0,80522	0,78281	0,78281
MNB	0,86520	0,84897	0,85147
SVM-linear	0,85640	0,85515	0,85515
SVM-poly	0,86392	**0,86392**	0,86267
SVM-rbf	0,86267	0,86019	0,86019
SemEval2015_Hotel	0,83333	**0,86667**	**0,86667**
C4.5-entropia	0,66667	**0,80000**	**0,80000**
C4.5-gini	0,66667	**0,80000**	**0,80000**
KNN-cosseno	0,80000	**0,86667**	**0,86667**
KNN-euclidiano	0,80000	**0,86667**	**0,86667**
MNB	0,76667	**0,83333**	**0,83333**
SVM-linear	0,80000	**0,80000**	**0,80000**
SVM-poly	0,83333	**0,83333**	**0,83333**
SVM-rbf	0,73333	**0,83333**	**0,83333**
SemEval2015_Laptop	0,88101	**0,88101**	**0,88101**
C4.5-entropia	0,73140	0,72442	0,72209
C4.5-gini	0,75255	0,73112	0,73112
KNN-cosseno	0,83184	0,78998	0,78765
KNN-euclidiano	0,83411	0,78522	0,78289
MNB	0,87386	0,85055	0,85055
SVM-linear	0,87868	**0,87868**	**0,87868**
SVM-poly	0,87403	0,87171	0,87171
SVM-rbf	0,88101	0,88101	0,88101
SemEval2015_Rest	0,88076	0,87790	0,87790
C4.5-entropia	0,83731	**0,83731**	**0,83731**
C4.5-gini	0,83739	0,78824	0,79983
KNN-cosseno	0,79950	**0,82261**	**0,83134**
KNN-euclidiano	0,79950	**0,82261**	**0,83134**
MNB	0,79950	**0,82261**	**0,83134**
SVM-linear	0,85765	**0,86059**	**0,86059**
SVM-poly	0,86042	**0,86622**	**0,86630**
SVM-rbf	0,83739	**0,85160**	**0,85445**

para SVM-linear, com valor de 0,84138. Para o conjunto de dados *SemEval2015_Hotel* os melhores valores foram para KNN (cosseno e euclideano), cujo valor foi de 0,86667 usando tanto gBoED_Freq quanto gBoED_Dist. Este último conjunto apresentou resultados bastante expressivos. Já *SemEval2015_Laptop* obteve desempenho bastante aquém com apenas um resultado superior a BOW.

Nos conjuntos de dados *SemEval2014* e *SemEval2015*, de maneira geral os melhores resultados, foram usando o método tradicional BOW. Acredita-se que a união de diferentes domínios possa prejudicar o desempenho do método, pois determinados termos podem possuir significados diferentes em domínios diferentes. Por exemplo,"*fila_grande*" e "*tela_grande*" pertencem a classes opostos em seus diferentes domínios.

Nos conjuntos de dados separados por domínio, tanto *SemEval2014_Laptop* quanto *SemEval2014_Rest* obtiveram um desempenho bastante semelhante ao conjunto completo. *SemEval2015_Rest* e *SemEval2015_Hotel* obtiveram resultados superiores ao conjunto completo. Destaque para o último conjunto, que apresentou resultados expressivos. Isso se deve ao fato do conjunto de documentos ser bastante restrito. Já *SemEval2015_Laptop* obteve desempenho bastante aquém com apenas um resultado superior a BOW. Comparativamente, em 3.284 casos, gBoED_Dist obteve maior acurácia do que o gBoED_Freq. Assim, o esquema de ponderação baseado na distância entre os termos ainda apresenta um impacto positivo na efetividade da gBoED.

4. Conclusões

Neste trabalho foi realizada a extensão dos experimentos de avaliação do método proposto por [Scheicher et al. 2019], que aplica expressões de domínio para melhorar a classificação de documentos com baixa confiança preditiva. A avaliação experimental extendida reforça a conclusão de que o método é adequado quando as revisões se referem a entidades da mesma natureza. Como trabalhos futuros, pretende-se: (i) aplicar o metódo proposto em um framework baseado em modelos de linguagem para validação do uso de expressões do domínio como informações enriquecidas, (ii) construção das expressões do domínio baseada na sintaxe, visando maior precisão nas expressões formadas e (iii) adaptar o método para outros domínios além da análise de sentimentos.

Agradecimentos

Agradecimentos aos auxílios fornecidos pela CAPES e FAPESP, processo 2016/17078-0.

Referências

Dosilovic, F. K., Brcic, M., and Hlupic, N. (2018). Explainable artificial intelligence: A survey. In *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215.

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3):50–57.

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

Ju, H. and Yu, H. (2018). Sentiment Classification with Convolutional Neural Network using Multiple Word Representations. In *12th Int. Conf. on Ubiquitous Information Management and Communication*, pages 1–7.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Pontiki, M., Galanis, D., Androutsopoulos, I., Manandhar, S., and Papageorgiou, H. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *8th International Workshop on Semantic Evaluation*, pages 27–35.

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

Scheicher, R. B., Sinoara, R. A., Felinto, J. C., and Rezende, S. O. (2019). Sentiment classification improvement using semantically enriched information. In *19th ACM Symposium on Document Engineering*, pages 1–4.

Scheicher, R. B., Sinoara, R. A., Koga, N. J., and Rezende, S. O. (2016). Uso de expressões do domínio na classificação automática de documentos. In *XIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 625 – 636.

Xiong, S. (2016). Improving twitter sentiment classification via multi-level sentiment-enriched word embeddings.