Existem domínios melhores que outros em Cross-Domain Sentiment Analysis?

Valéria Pereira de Souza
Universidade Federal de Minas Gerais
valeriaps@ufmg.br

RESUMO

O cross-domain sentiment analysis estuda a tarefa de análise de sentimento na situação onde o domínio a ser predito é diferente do domínio no qual o modelo de machine learning foi treinado. Esse cenário apresenta muitos desafios aos pesquisadores e as principais estratégias envolvem a utilização parcial de um dataset do domínio alvo, mesmo que não anotado.

Esse estudo investiga estratégias para realizar análises de sentimento dada a ausência completa de informações prévias do domínio alvo. Para isso, uma rede CNN é treinada em cinco datasets diferentes com avaliações de produtos da Amazon. Foram experimentadas três estratégias: o uso de um único domínio, a combinação completa de todos os domínios, com exceção do domínio alvo e, por último, a escolha orientada de um domínio para complementar o treinamento.

As estratégias de combinação de domínios mostrou-se infrutífera ao tempo em que foi possível identificar que o domínio fonte 'computers & videogames' produz resultados médios melhores do que todos os demais e melhores do que as estratégias de combinação de datasets.

PALAVRAS CHAVE

cross-domain; sentiment analysis; neural networks.

1 INTRODUÇÃO

A tarefa de análise de sentimento é considerada bem resolvida pela comunidade científica de processamento de linguagem natural. As abordagens estado da arte envolvem treinar o modelo em dataset anotado do mesmo domínio que se quer predizer. Entretanto, na aplicação cotidiana da tarefa, nem sempre datasets anotados estão disponíveis para o domínio que se quer predizer - domínio alvo, e, ao se treinar um modelo em determinado domínio, seu poder preditivo vai apresentar resultados piores se utilizado para predizer um domínio diferente.

A frente de estudos em cross-domain sentiment analysis investiga técnicas para melhorar os resultados quando da ausência de um dataset anotado para o domínio alvo a partir de datasets anotados de outros domínios - domínios fonte. Algumas técnicas envolvem a anotação de uma pequena parte do dataset do domínio-alvo para complementar o treinamento do modelo. Outras técnicas envolvem utilizar o dataset do domínio alvo mesmo sem anotação prévia para enriquecimento semântico do corpus já disponível [1].

1.1 Desafios de cross-domain sentiment analysis

A queda no desempenho entre avaliações in-domain e cross-domain é bem documentada pela literatura e ocorre devido a quatro principais fatores: esparsidade, polissemia, feature divergence e polarity divergence[2]. A esparsidade ocorre quando o domínio alvo contém palavras não existentes no domínio fonte. A polissemia é a existência de mais de um significado para uma mesma palavra. Feature divergence é a existência de palavras muito específicas ao domínio fonte, que não aparecem no domínio alvo.

Por último, polarity divergence é a possibilidade de uma mesma feature representar sentimentos opostos em domínios diferentes. A palavra 'hot', por exemplo, pode ser negativo, se em uma avaliação de celular, mas positiva, se em uma avaliação de um jogo.

2 METODOLOGIA

2.1 Datasets

Foi utilizado o conjunto de datasets anotados Amazon multi domain[1] formado por avaliações de produtos adquiridos na plataforma. O conjunto apresenta 25 datasets de domínios diferentes possibilitando muita experimentação. Foram selecionados cinco domínios de natureza bem diversa entre si com o objetivo de dificultar o trabalho do modelo: 'apparel', 'baby', 'computer & videogames' (comp & games), 'grocery' e 'sports'.

2.2 Pré processamento

Após o parsing do xml, as palavras foram colocadas em minúsculo e foram removidas as stop words. Na sequências, foram removidas palavras que aparecem em menos do que 5 documentos por corpus. Datas e números foram substituídos pelos termos "parsedyear" e "parseddigits". A lematização foi feita com o WordNetLemmatizer, da biblioteca NLTK.

Para trabalhar com classificação binária, as avaliações 1 e 2 foram convertidas em 0 (negativas) e as 4 e 5, em 1 (positivas). Foram ignoradas as avaliações com nota 3.

O arquivo original apresenta as avaliações em ordem crescente de nota, o que poderia representar um problema na divisão em base de treino e base de teste, desbalanceando a proporção de labels entre os arquivos. Dessa forma, as linhas do arquivo tratado foram embaralhadas.

O tamanho do vocabulário foi padronizado nas 10000 palavras mais frequentes para cada domínio, com exceção de 'grocery', que apresentou feature divergence: após pré-processamento, continha 7490 palavras, de forma que, para esse domínio, esse foi o tamanho considerado.

2.3 Arquitetura

O modelo utilizado foi uma rede neural do tipo CNN com MaxPooling, com embedding word2vec na arquitetura skip-gram. A implementação foi realizada com a biblioteca Keras.

[1] Disponível em <https://www.cs.jhu.edu/~mdredze/datasets/sentiment/>

3 RESULTADOS

3.1 Baseline

Para visualizar a variação na performance, bem como gerar baselines, o modelo foi treinado em cada domínio e avaliado nos demais. A tabela 1 apresenta a comparação entre a avaliação in-domain e o melhor resultado nas avaliações cross-domain (baseline).

Tabela 1: comparação entre a avaliação in-domain e o melhor resultado cross-domain

Alvo	In domain	Baseline		Diferença (%)
		Fonte	Acurácia	
Apparel	93,13	Sports	81,80	*-12,33*
Baby	89,31	Grocery	79,09	*-11,44*
Comp & games	96,93	Grocery	86,88	*-10,36*
Grocery	96,54	Comp &games	83,57	*-13,43*
Sports	90,13	Grocery	80,80	*-10,35*

Para todos os casos, houve redução do desempenho, como esperado. Os domínios fonte que geraram os melhores resultados, entretanto, não são óbvios. Ainda que seja possível alegar que a categoria 'sports' vende muitas roupas, e por isso uma melhor capacidade preditiva para 'apparel', seria muito mais difícil identificar argumentos para a dupla baby-grocery.

3.2 Poder de predição dos domínios

Se é fato que cada domínio apresenta peculiaridades, também apresentam diferentes graus de generalidade. A tabela 2 2 apresenta um resumo das acurácias para os domínios estudados. O domínio 'baby' apresentou o pior desempenho geral como domínio fonte, enquanto 'comp & games', o melhor, mesmo que 'grocery' seja a moda como baseline.

A tabela 3 apresenta, para cada domínio-alvo, a acurácia obtida com diferentes domínios fonte; uma espécie de medida de "preditibilidade". O domínio 'comp & games'', enquanto domínio alvo, aparentemente é o mais sensível a trocas de domínio ao tempo em que foi o melhor preditor. O domínio-alvo 'grocery' apresentou a menor

diferença entre a acurácia média e o baseline, aceitando melhor a variação de domínios fonte.

Tabela 2: acurácia geral por domínio fonte

Fonte	Acurácia média
Apparel	*77,22 ± 4,09*
Baby	***71,69 ± 5,74***
Comp & games	***81,61 ± 4,35***
Grocery	*77,95 ± 4,09*
Sports	*78,40 ± 4,40*

Tabela 3: acurácia geral por domínio alvo

Alvo	Acurácia média	Diferença para o baseline (%)
Apparel	*76,71 ± 5,75*	*6,22*
Baby	***73,33 ± 3,26***	*7,28*
Comp & games	*78,48 ± 3,88*	***9,66***
Grocery	***81,60 ± 2,68***	***2,35***
Sports	*75,00 ± 4,26*	*7,17*

Aparentemente 'baby' é o domínio que apresenta mais particularidades em sua semântica e features não conseguindo ser bem predito nem predizer outros domínios.

3.1 União dos domínios fonte

O primeiro experimento para verificação do que ocorre com a acurácia foi a junção de todos os domínios fonte, com exceção do domínio alvo, resultado demonstrado na tabela 4.

Para todos os casos, a acurácia apresentou piora. Em alguns casos, a piora com relação ao baseline foi maior do que a piora do baseline com relação à avaliação in-domain. A hipótese de que ao "ver" mais usos da palavra, o modelo fosse capaz de generalizar para mais contextos, não se sustentou; aparentemente o excesso de possíveis contextos e semânticas dilui a capacidade preditiva do modelo.

3.1 Escolha de uma fonte complementar

O segundo experimento foi a seleção de um único outro dataset para complementar o treinamento da fonte para as piores duplas, para cada alvo. A fonte complementar selecionada foi o domínio fonte baseline para aquele alvo.

Tabela 4: Resultado para o treinamento com união de todos os domínios fonte, com exceção do domínio alvo

Alvo	Acurácia	Diferença (%)
Apparel	78,42	***-4,13***
Baby	67,80	***-14,27***
Comp &games	74,01	***-14,81***
Grocery	80,04	***-4,22***
Sports	77,82	***-3,69***

O foco foi investigar se reduzindo o excesso de variação de contextos (em relação à união completa das fontes), era possível melhorar o resultado. O resultado da dupla foi melhorado em quatro dos cinco casos, entretanto, em nenhum caso a acurácia foi superior ao uso do melhor domínio-fonte.

Tabela 5: resultados da adição de uma fonte complementar

Fonte-Alvo	Acc.	Com fonte complementar	
		Acc.	Dif. para o baseline(%)
Sports - Apparel	**72,66**	**78,69**	***-2,61***
Apparel - Baby	**66,52**	**74,36**	***-9,09***
Baby - Comp &games	77,56	72,97	***-7,73***
Apparel - Grocery	**74,69**	**74,96**	***-8,36***
Baby - Sports	**71,68**	**74,15**	***-6,24***

4 CONCLUSÃO

A tarefa de análise de sentimento na situação de cross-domain apresenta desafios não triviais. A completa ausência de informações do domínio alvo significa uma redução significativa na acurácia. A formação de datasets a partir de mais de um domínio não se mostrou produtiva frente ao uso do melhor domínio-fonte.

Por outro lado, em uma situação de ausência completa de informações do domínio alvo também

não seria possível identificar qual o domínio fonte mais adequado. Para os datasets investigados, o domínio 'comp & games' parece ser o que produz os melhores resultados, em média, inclusive superando os experimentos com combinações de domínios.

A noção de domínios parecidos ou diferentes, baseada em senso comum, e utilizada para seleção dos datasets, pode ser enganosa e não é válida como boa fonte de seleção de domínios.

Futuramente pode ser interessante investigar até onde se entende o poder preditivo do domínio 'comp & games' e a dificuldade de generalização do domínio 'baby' com experimentos em mais domínios de produtos.

4 REFERÊNCIAS BIBLIOGRÁFICAS

[1] Hasegawa, Marcello, and Praveen Rokkam. "Domain specific sentiment analysis using cross-domain data."

[2] Al-Moslmi, Tareq, et al. "Approaches to cross-domain sentiment analysis: a systematic literature review." *IEEE Access* 5 (2017): 16173-16192.

5. ANEXO

Fonte	Alvo	Acurácia
Apparel Acurácia in domain: 93,31	Computer	83,67
	Sports	80,26
	Grocery	74,69
	Baby	66,52
Baby Acurácia in domain: 89,31	Computer	77,56
	Grocery	74,85
	Apparel	74,74
	Sports	71,68
Comp&games Acurácia in domain: 96,93	Grocery	83,57
	Sports	80,87
	Apparel	80,32
	Baby	73,65
Grocery Acurácia in domain: 96,54	Computer	86,88
	Apparel	81,18
	Sports	80,80
	Baby	79,09
Sports Acurácia in domain: 90,13	Grocery	78,69
	Computer	78,34
	Apparel	72,66
	Baby	67,51