Classificação de subjetividade para a língua portuguesa

Luana Balador Belisário, Luiz Gabriel Ferreira, Thiago Alexandre Salgueiro Pardo
Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos/SP, Brasil
www.nilc.icmc.usp.br
luana.belisario@usp.br, luizgferreira@usp.br, taspardo@icmc.usp.br

Keywords **-** análise de sentimentos, classificação de subjetividade, léxico, aprendizado de máquina

I. Introdução

Textos publicados nas redes sociais têm sido uma fonte valiosa de informações para organizações, uma vez que a análise desses textos é uma forma de aprimorar o feedback de produtos e serviços dessas empresas.

Devido ao crescimento do número de usuários em redes sociais e plataformas web, tornou-se possível obter grande volume de dados para esse tipo de análise. Com isso, surgiu a necessidade da criação de ferramentas que pudessem analisar as opiniões dos usuários de forma automática. A área de pesquisa que realiza esse tipo de processamento é a Análise de Sentimentos, também chamada de Mineração de Opiniões.

A análise de subjetividade é uma das primeiras etapas na mineração de opiniões. Nessa tarefa, os documentos de interesse (que podem ser textos completos, sentenças ou mesmo fragmentos menores) são classificados como subjetivos ou objetivos de acordo com sua polaridade: quando classificados como objetivos, assume-se que expressam fatos; quando ditos subjetivos, expressam opiniões e sentimentos (Liu, 2012).

Fatos apontam informações sobre um acontecimento (por exemplo, "Comprei um netbook Philco em nov/2010."). Opiniões, por sua vez, expressam o ponto de vista de uma ou mais pessoas a respeito de um tema (por exemplo, "Esse livro que lí é muito bom, de uma profundidade incrível!"). No exemplo factual, é possível identificar sua objetividade por se tratar de uma notícia e por não ser possível identificar se a notícia é boa ou ruim por princípio. Já no exemplo opinativo, é possível identificar que o autor do comentário gostou no livro lido através das expressões "muito bom" e "incrível". Essas palavras que denotam evidentemente a opinião e a polaridade dela são chamadas de palavras de sentimento.

Na Tabela 1, pode-se observar algumas outras sentenças classificadas que ocorreram no córpus Computer-BR utilizado pelos autores Moraes et al. (2016). As sentenças subjetivas são ainda subdivididas em polaridade "positiva" e "negativa", expressando explicitamente sentimento em relação a um notebook, com indicativos como "tô mt feliz" e "Parece brincadeira de mau gosto". As sentenças objetivas, apesar de possuírem adjetivos que podem indicar opinião como "boas" e "baum" (bom), não expressam opinião a respeito dos produtos. Cabe destacar ainda a linguagem utilizada, que apresenta abreviações e vocabulário típico do domínio. Essas são considerações que devem ser feitas ao desenvolver métodos para este tipo de classificação.

Tabela 1: Exemplos de sentenças rotuladas do córpus Computer-BR.

Sentença	Polaridade
Sem notebook de novo... Parece brincadeira de mau gosto	Subjetiva/Negativa
Alguem me indica marcas boas de notebook?	Objetiva
Logo um precioso notebook Dell lindíssimo chega aqui em casa tô mt feliz	Subjetiva/Positiva
Esse Not é baum ? Alguem sabe ?	Objetiva

Para a língua portuguesa, há muitos trabalhos na área de análise de sentimentos. Entretanto, no melhor do nosso conhecimento, há apenas uma iniciativa (Moraes et al., 2016) dedicada à tarefa de classificação de subjetividade. Essa tarefa é muito relevante para etapas posteriores de análise textual, sendo responsável por indicar o conteúdo relevante para processamento, ou seja, as sentenças subjetivas, que são de maior interesse em mineração de opiniões. Essa seleção de conteúdo tem potencial para aprimorar os resultados de aplicações relacionadas na área.

Neste artigo, iniciamos por reproduzir os experimentos de Moraes et al. Além disso, estendemos a avaliação dos métodos para outros córpus, visando avaliar sua robustez. Especificamente, comparamos abordagens para classificação de subjetividade baseada em léxico e baseada em aprendizado de máquina. Mostramos, ao final, que o melhor resultado obtido foi de 77% de acurácia no método baseado em aprendizado de máquina e de 75,2% no método baseado em léxico. Além disso, evidenciamos que fatores como tamanho do córpus, balanceamento entre sentenças de diferentes polaridades e diferentes técnicas de pré-processamento aplicadas nos córpus influenciam significativamente os resultados.

Na seção seguinte, fazemos uma breve revisão literária. Em seguida, apresentamos os córpus utilizados neste trabalho. Nas Seções IV e V, descrevemos os métodos avaliados. Os resultados principais são relatados na Seção VI. Por fim, fazemos algumas considerações finais na Seção VII.

II. Revisão Bibliográfica

Moraes et al. (2016) propõem métodos para efetivamente classificar a subjetividade de textos em nível sentencial para a língua portuguesa. Os autores argumentam sobre a importância dessa área de pesquisa para empresas no que diz respeito a um feedback mais completo e eficiente de produtos ou serviços oferecidos por elas.

Para testar os métodos, os autores do artigo criaram um córpus com 2.317 tweets sobre a área de tecnologia. Esse córpus - chamado Computer-BR - foi anotado manualmente e submetido a um pré-processamento para aumentar a eficiência dos métodos aplicados. Detalhes sobre o córpus e o pré-processamento estarão na próxima seção.

Os métodos abordados e testados pelos autores são baseados em léxico e em aprendizado de máquina. Os melhores resultados com os métodos baseados em léxico atingiram 64% de medida-f, enquanto os resultados dos métodos baseados em aprendizado de máquina chegaram a 75%. Os métodos, implementados neste artigo, são descritos posteriormente.

III. Córpus Utilizados

Neste trabalho, para o teste dos dois métodos, foram usados três córpus compostos por revisões de produtos ou serviços de três domínios distintos.

Como especificado na seção anterior, o córpus Computer-BR contém 2.317 sentenças com mensagens de usuários sobre a área de tecnologia. É interessante observar que esse córpus apresenta desbalanceamento das classes, com quantidade superior de sentenças objetivas.

Além do domínio de tecnologia no córpus Computer-BR, utilizamos um córpus de resenhas de livros, que reúne resenhas curtas de leitores retiradas do córpus ReLi (Freitas et al., 2012), do site de compras Amazon e da rede social Skoob, com 270 sentenças divididas igualmente entre fatos e opiniões. Essas sentenças foram classificadas manualmente.

Por fim, utilizamos um córpus de produtos eletrônicos com 230 sentenças que foram retiradas do conhecido córpus Buscapé (Hartmann et al., 2014) e anotadas posteriormente de forma manual. Assim como o Computer-BR, esse córpus apresenta desbalanceamento, com cerca de 70% de sentenças classificadas como subjetivas.

Os três córpus utilizados são formados por sentenças retiradas da web. Logo, o grande número de abreviações, erros de ortografia e jargões específicos caracterizam esse meio e comprometem o desempenho dos métodos automáticos de classificação. Para amenizar essa dificuldade foram realizadas algumas etapas de pré-processamento. Primeiramente, as sentenças passaram por um processo de normalização da linguagem com o auxílio do sistema de normalização textual *enelvo*[1] (Bertaglia, 2017). Após isso, as *stopwords* (conforme constam no pacote *Natural Language Toolkit*[2] - NLTK), a pontuação, os números e os demais caracteres especiais foram removidos. Por fim, os termos foram reduzidos às suas formas canônicas com o auxílio de um lematizador[3] desenvolvido no NILC.

A seguir, apresentamos os métodos implementados e avaliados neste artigo.

IV. Método baseado em léxico

O artigo de Moraes et al. abordou três heurísticas para classificação de subjetividade. As três consistem em pesquisar palavra por palavra da sentença de interesse e verificar a subjetividade e polaridade de cada uma. A Heurística 1 se traduz em somar a polaridade de todas as palavras subjetivas e o resultado já é a classificação; a Heurística 2 classifica uma sentença em subjetiva se ela possui um valor mínimo de palavras subjetivas, e então é feito o cálculo para saber sua polaridade; já na Heurística 3, a sentença é classificada como subjetiva se ela possui uma proporção mínima de palavras subjetivas, ou seja, quantas palavras do total de palavras da sentença são subjetivas. A Heurística 1 foi a que produziu os melhores resultados no artigo de Moraes et al.

A Heurística 1 depende de um léxico de palavras subjetivas (ou palavras de sentimento) pré-classificadas para analisar o texto em nível sentencial. Essas palavras foram associadas com valor 1 se eram positivas, -1 para negativas e 0 para neutras. Basicamente, a heurística consiste em determinar subjetividade da sentença somando as polaridades das palavras que a compõem. A fórmula está representada em (1), onde n é o número de palavras da sentença, 'term i' representa cada palavra da sentença e subjetividade(sent) é a subjetividade em nível sentencial:

$$\textbf{subjetividade(sent)} = \sum_{i=1}^{n} \textbf{polaridade(term i)} \quad (1)$$

Se a subjetividade(sent) assumir um valor diferente de 0, a sentença é considerada subjetiva, sendo valor positivo (maior que 0) uma sentença "subjetiva positiva" e valor negativo (menor que 0) uma sentença "subjetiva negativa". Caso o valor seja zero, a sentença é considerada "objetiva" ou "neutra". Em razão da simplicidade do método, não há tratamento de negação, ironia e advérbios, cujas funções seriam intensificar, neutralizar ou até mudar a orientação das palavras de sentimento.

V. Método baseado em Aprendizado de Máquina

Para as técnicas baseadas em aprendizado de máquina, foram testados dois algoritmos de classificação, seguindo-se a proposta de Moraes et al.. Ambos utilizam o modelo *bag*

[1] https://github.com/tfcbertaglia/enelvo
[2] https://www.nltk.org/
[3] http://conteudo.icmc.usp.br/pessoas/taspardo/LematizadorV2a.rar

of words, que seleciona cada palavra contida na sentença como um atributo distinto. As palavras que poderão ser utilizadas como atributos foram limitadas com base em suas relevâncias para melhor desempenho. A seguir são descritos os métodos de seleção das palavras.

A etapa inicial consiste em quantificar a relevância das palavras em cada classe. Para isso foram utilizadas duas métricas, sendo a primeira a frequência da palavra na classe, como mostrado em (2), onde w_k indica a palavra e c_j indica a classe.

$$\mathbf{freq = P(w_k \mid c_j)} \quad (2)$$

A segunda métrica utilizou o *Comprehensive Measurement Feature Selection* (CMFS), que busca calcular a relevância da palavra na classe considerando as suas ocorrências nas outras classes. Para isso, realiza-se o produto da probabilidade da palavra pertencer à classe pela sua frequência na classe:

$$\mathbf{CMFS(w_k, c_j) = P(w_k \mid c_j).P(c_j \mid w_k)} \quad (3)$$

Com as métricas descritas, é gerada uma lista para cada classe com as palavras mais relevantes. A partir dessas listas, é gerada uma única lista com as palavras que serão utilizadas como atributos. De cada lista, são selecionadas as *n* (*n* menor que 100) palavras mais relevantes e agrupadas de duas maneiras distintas. A primeira simplesmente junta todas as palavras, e a segunda junta as palavras, mas exclui as que aparecem nas duas listas. Em ambos os casos, caso se obtenha o mesmo valor para mais de uma palavra na última posição, é utilizada apenas uma.

Os algoritmos utilizados são o *Naive Bayes* (NB) e o *Sequential Minimal Optimization* (SMO). O primeiro é baseado no teorema de Bayes e calcula previamente as probabilidades dos termos nas classes. Para realizar a classificação, assume-se que os atributos sejam condicionalmente independentes e escolhe-se a classe que tenha a maior probabilidade. O SMO é uma otimização matemática para treinar as "máquinas de vetores de suporte". O objetivo do método é dividir o espaço descrito pelos atributos em duas regiões, cada uma pertencente a uma classe. Na classificação, analisa-se em qual região do espaço o conjunto de termos da sentença se situa.

VI. Resultados e Discussão

Os resultados alcançados para as classes "subjetivo" e "objetivo" são mostrados nas Tabelas 2 e 3, respectivamente. São exibidos valores para as medidas clássicas de precisão, cobertura e medida-f, além da acurácia geral. Exibimos apenas os melhores resultados conseguidos.

Como se pode notar nas tabelas, para os métodos baseados em léxico, fizemos experimentos com os léxicos de sentimento WordnetAffectBR (Pasqualotti e Vieira, 2008) e Sentilex-PT (Carvalho e Silva, 2015), amplamente conhecidos na área.

Tabela 2: Melhores resultados em abordagens baseadas em léxico e aprendizado de máquina para cada córpus para sentenças subjetivas.

Córpus	Computer-BR		Resenha de Livros		Produtos Eletrônicos	
Método	NB	WordnetAffect	SMO	WordnetAffect	SMO	Sentilex
Precisão	0.580	0.524	0.817	0.736	0.906	0.912
Cobertura	0.630	0.189	0.703	0.293	0.694	0.768
Medida-f	0.601	0.278	0.744	0.419	0.782	0.834
Acurácia	0.771	0.729	0.770	0.600	0.700	0.752

Tabela 3: Melhores resultados em abordagens baseadas em léxico e AM para cada córpus para sentenças objetivas.

Córpus	Computer-BR		Resenha de Livros		Produtos Eletrônicos	
Método	NB	WordnetAffect	SMO	WordnetAffect	SMO	Sentilex
Precisão	0.853	0.751	0.734	0.567	0.373	0.405
Cobertura	0.825	0.934	0.836	0.898	0.729	0.682
Medida-f	0.838	0.833	0.770	0.695	0.475	0.501
Acurácia	0.771	0.729	0.770	0.600	0.700	0.752

Analisando-se as tabelas, é possível inferir que os resultados para o córpus Computer-BR ficaram próximos dos descritos por Moraes et al. (2016), com variações decorrentes, principalmente, das técnicas diferentes de pré-processamento. Para a acurácia, por exemplo, os melhores valores originalmente eram 0.78 para técnicas de aprendizado de máquina e 0.74 para abordagens baseadas em léxico, enquanto os valores obtidos aqui foram 0.77 e 0.73, respectivamente.

É perceptível que o córpus de produtos eletrônicos apresentou piores resultados em relação ao córpus de resenha de livros, principalmente ao utilizar métodos baseados em aprendizado de máquina, evidenciada na classificação de sentenças objetivas. Essa piora pode ser justificada pelo desbalanceamento do córpus, fato que ocorre também no Computer-BR. Além disso, ao comparar com o Computer-BR, que possui aproximadamente dez vezes o número de sentenças, nota-se a influência do tamanho do córpus para o aprendizado de máquina.

Desconsiderando o desbalanceamento, os erros cometidos pelos algoritmos de aprendizado de máquina decorrem, principalmente, da falta de informações sobre alguns termos. Esse fato ocorre com mais frequência nos córpus de avaliação de produtos eletrônicos e Computer-BR, onde a linguagem é mais informal, gerando baixa frequência para diversos termos que possuem significados similares. Como exemplo, temos a sentença retirada do Computer-BR a seguir: "Notebook da Positivo é péssimo, credo! Melhor notebook é Dell". Os termos "péssimo", "credo" e "melhor" caracterizam a polaridade da sentença, tornando simples uma classificação que tenha informação semântica a respeito. Entretanto, os termos aparecem poucas vezes no córpus e acabam não sendo selecionados como atributos, o que deixa termos que não possuem polaridade, como "notebook", "dell" e "positivo", como possíveis atributos em uma classificação.

Outra limitação do método é por conta do tratamento dos termos fora de contexto. Apesar de ser funcional em casos onde é possível extrair termos que possuem polaridade nítida, em outros casos a polaridade pode ser definida pela

maneira como a sentença foi construída, e o método não é capaz de distinguir.

Em relação ao método baseado em léxico, por se fundamentar simplesmente na busca e contagem das palavras de sentimento, não se tratam com eficiência possíveis desvios na polaridade das sentenças, como o tratamento de sarcasmo e ironia, advérbios de negação e intensidade e desambiguação de palavras. Um exemplo dos problemas citados ocorreria ao classificar a sentença "O processador desse notebook é pouco eficiente". Nesse caso, o algoritmo encontraria a única palavra de sentimento que seria "eficiente" e classificaria como subjetiva e positiva a sentença, porém, sabe-se que um processador ser pouco eficiente é uma característica negativa para o produto. O advérbio de intensidade "pouco" alterou a polaridade da palavra de sentimento "eficiente", mas isso não foi detectado pelo método.

Por fim, destaca-se a grande variação de resultados entre os córpus diferentes, corroborando o que se observa na literatura da área. Em particular, para o método baseado em léxico, o léxico WordnetAffectBR foi mais apropriado em dois dos três domínios. Em aprendizado de máquina, o método SMO se destacou em relação ao NB.

A seguir, apresentamos algumas considerações finais.

VII. Considerações finais

O trabalho descrito apresentou duas abordagens para a classificação automática de sentenças retiradas da web, tendo como objetivo a distinção entre sentenças subjetivas e objetivas. Os métodos apresentados constituem uma reprodução dos utilizados originalmente por Moraes et al. (2016), estendidos para outros córpus. Apesar dos resultados promissores, ainda há muito a se fazer.

Em aprendizado de máquina, visa-se ainda investigar e desenvolver métodos com o uso de *word embeddings* (Mikolov et al., 2013). Na abordagem baseada em léxico, o próximo passo será implementar o método de classificação de subjetividade baseado em grafos abordado por Vilarinho et al (2018). Há, também, desafios relacionados às classes das sentenças. Por exemplo, é sabido que sentenças objetivas podem carregar opinião. Entretanto, esse fenômeno tem sido deixado de lado nesta pesquisa, podendo ser investigado em trabalhos futuros.

Por fim, observa-se que este trabalho integra o projeto maior OPINANDO (*Opinion Mining for Portuguese: Concept-based Approaches and Beyond*)[4], em cuja página encontram-se vários dos recursos e ferramentas citados aqui.

Agradecimentos

À Fundação de Amparo à Pesquisa do Estado de São Paulo (processo nro 2018/11479-9) e à Pró-Reitoria de Pesquisa da USP pelo apoio a este projeto.

[4] https://sites.google.com/icmc.usp.br/opinando/

Referências

Bertaglia, T.F.P. (2017). Normalização textual de conteúdo gerado por usuário. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. 160p.

Carvalho, P. e Silva, M.J. (2015). Sentilex-PT: Principais Características e Potencialidades. Oslo Studies in Language, Vol. 7, N. 1, pp. 425-438.

Freitas, C.; Motta, E.; Milidiú, R.; Cesar, J. (2012). Vampiro que brilha... rá! Desafios na anotação de opinião em um corpus de resenhas de livros. In the Proceedings of the XI Encontro de Linguística de Corpus (ELC), pp. 1-12.

Hartmann, N. S.; Avanço. L.; Balage, P. P.; Duran, M. S.; Nunes, M. G. V.; Pardo, T.; Aluísio, S. (2014). A Large Opinion Corpus in Portuguese - Tackling Out-Of-Vocabulary Words. In the Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC). pp. 3865-3871.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. 168p.

Mikolov, T.; Corrado, G.; Chen, K.; Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv: 1301.3781, pp. 1-12.

Moraes, S.M.W.; Santos, A.L.L.; Redecker, M.; Machado, R.M.; Meneguzzi, F.R. (2016). Comparing Approaches to Subjectivity Classification: A Study on Portuguese Tweets. In the Proceedings of the International Conference on the Computational Processing of the Portuguese Language (PROPOR), pp. 86-94

Pasqualotti, P.R. e Vieira, R. (2008). WordnetAffectBR: uma base lexical de palavras de emoções para a língua portuguesa. Revista Novas Tecnologias na Educação, Vol. 6, N. 2, pp. 1-10.

Vilarinho, G. N.; Ruiz, E. E. S. (2018). Global centrality measures in word graphs for Twitter sentiment analysis. In the Proceedings of the 7th Brazilian Conference on Intelligent Systems (BRACIS), pp. 55-60.