

Análise de Métodos de Extração de Aspectos em Opiniões Regulares

Title: Analysis of Aspect Extraction Methods in Regular Opinions

João Paulo Albuquerque Vieira¹, Raimundo Santos Moura¹

¹ Departamento de Computação
Universidade Federal do Piauí (UFPI)
Teresina, Piauí – Brasil
joaopauloalbu@gmail.com, rsm@ufpi.edu.br

Abstract. *This work presents a comparative analysis between the main approaches used at the task of Extraction of Aspects in reports about products and services on web sites. Adaptations of four methods of extraction of aspects were implemented and evaluated using two distinct Corpora, one in Portuguese and another in English. On the experiments performed it was observed that the approach using supervised learning (convolutional neural networks) obtained better results on the others.*

Keywords. *Opinion Mining; Sentiment Analysis; Aspect Extraction.*

Resumo. *Este trabalho apresenta uma análise comparativa entre as principais abordagens usadas na tarefa de Extração de Aspectos em comentários sobre produtos e serviços em web sites. Adaptações de quatro métodos de extração de aspectos foram implementadas e avaliadas usando dois Corpora distintos: um em português e outro em inglês. Nos experimentos realizados foi observado que a abordagem usando aprendizagem supervisionada (redes neurais convolucionais) obteve melhores resultados que as demais.*

Palavras-Chave. *Mineração de Opinião; Análise de Sentimentos; Extração de Aspectos.*

1. Introdução

Nos últimos anos, o comportamento dos usuários na Web vem mudando, pois, além de consumir os conteúdos disponíveis, eles estão cada vez mais fazendo compras online e compartilhando suas opiniões e experiências sobre os produtos e serviços. Um estudo encomendado pela Cisco IBSG revela que quase oito em cada dez consumidores são compradores digitais que pesquisam e compram produtos regularmente usando a Web. A pesquisa ainda mostra que as três fontes de informações mais importantes usadas para ajudar a tomar decisões de compra foram as avaliações e relatos nos sites varejistas (52%), as recomendações de amigos e familiares (49%) e os *reviews* de especialistas (42%) [Fretwell et al. 2013].

Cui et al. (2014) analisando os efeitos dos relatos de consumidores nas vendas de produtos eletrônicos revela que estes têm um efeito significativo nas vendas de novos

produtos e que a quantidade de críticas negativas tem um efeito maior do que as críticas positivas, confirmando o viés de negatividade. Bickart & Schindler (2001) ainda apontam que os relatos constituem uma importante fonte de informação adicional que, frequentemente, não está disponível na descrição do produto ou serviço. Além disso, as informações providas por indivíduos na Web demonstram ser mais confiáveis que as informações fornecidas pelos vendedores.

Devido ao grande volume de dados produzidos na Web, pesquisadores da área de Processamento de Linguagem Natural (PLN) têm buscado extrair informações úteis de dados não estruturados, pois é dessa forma que cerca de 95% das informações relevantes são produzidas¹. Esta enorme quantidade de dados possibilita a ampliação das fontes de opiniões quantitativa e qualitativamente, tornando as formas de coleta mais baratas e reduzindo o tempo necessário para obtenção de informações. Portanto, as empresas não precisam mais sair em busca da opinião dos seus clientes em dispendiosas pesquisas de campo, gastando recursos para ter o *feedback* necessário para melhorar os seus investimentos e, conseqüentemente, aumentar os lucros [Liu 2015].

Como a quantidade de dados disponíveis para análise excede a capacidade humana para uma investigação manual, torna-se necessário o uso de métodos e técnicas para processar automaticamente não apenas o conteúdo dos relatos, mas também a opinião e o sentimento que expressam, conseguindo extrair informações realmente úteis sobre os mesmos para aquisição de novos conhecimentos [Cambria 2016, Barbosa et al. 2016].

A oportunidade de capturar a opinião de um público geral tem levado ao crescente interesse da comunidade científica, por causa dos grandes desafios abertos e da comunidade de negócios, em razão dos notáveis benefícios para a área de *marketing* e previsões do mercado financeiro. Mais recentemente, com a evolução do aprendizado de máquina profundo, a Análise de Sentimentos tem recebido mais atenção e vigor ao avançar sobre várias tarefas [Young et al. 2018, Cambria et al. 2017, Cambria 2016, Barbosa et al. 2016].

Formalmente, Análise de Sentimentos (do inglês, *Sentiment Analysis*)², é a subárea de estudo que analisa as opiniões que as pessoas expressam sobre um determinado assunto em textos escritos em linguagem natural. Ela é estruturada em três etapas: i) identificação das opiniões expressas sobre determinado assunto ou alvo em um conjunto de documentos; ii) classificação da orientação semântica ou polaridade da opinião (se tende a ser positiva, negativa ou neutra); e iii) apresentação dos resultados de forma agregada e sumarizada.

Na etapa de identificação, o processo de extração de aspectos é a principal tarefa e está relacionada à descoberta de conhecimentos por meio de técnicas de análise e extração de dados a partir de textos livres. Aspectos são atributos ou componentes de entidades, por exemplo, a entidade *Galaxy S6* tem aspectos como *custo benefício*, *durabilidade*, *bateria*, entre outros [Zhang & Liu 2014]. Este processo envolve a aplicação de algoritmos computacionais para analisar descrições textuais na busca de informações úteis e importantes. Na literatura especializada, existem algumas abordagens

¹ Disponível em <https://www.clarabridge.com/nlp-natural-language-processing/>

² Destaca-se que os termos Análise de Sentimentos e Mineração de Opinião foram utilizados como sinônimos em algumas publicações da área [Liu 2012]. No entanto, em trabalhos mais recentes tem-se preferido utilizar apenas Análise de Sentimentos [Liu 2015].

que buscam solucionar esta tarefa e podem ser agrupadas em quatro categorias [Liu 2012]: i) extração baseada em frequência; ii) extração baseada nas relações sintáticas; iii) extração usando aprendizado supervisionado; e iv) extração usando modelos de tópicos.

Adicionalmente, a Análise de Sentimentos Baseada em Aspectos é analisada na literatura por diversas tarefas diferentes, observando diversas perspectivas e usando diversos *Corpora*. Essas variações têm dificultado a comparação entre os métodos usados em sistemas de Análise de Sentimentos [Pontiki et al. 2014]. Para minimizar este problema, este artigo implementa um método para cada uma das quatro categorias existentes e realiza uma análise comparativa com os resultados obtidos, observando as mesmas perspectivas e os mesmos *Corpora*.

Portanto, o objetivo geral é fazer uma análise comparativa entre métodos das quatro abordagens para extrair aspectos em textos opinativos no domínio de produtos e serviços. A nossa intenção é avaliar os métodos implementados e descobrir o comportamento das abordagens em *Corpora* de relatos nos idiomas português e inglês. São consideradas contribuições deste trabalho: i) criação de um *Corpus* em língua portuguesa de relatos de consumidores contendo aspectos anotados; ii) implementação de métodos para extração baseado em frequência, relações sintáticas, aprendizado supervisionado e modelos de tópicos; e iii) análise dos métodos implementados e discussão do comportamento entre os *Corpora*.

Além desta seção introdutória, este trabalho está organizado da seguinte maneira: a Seção 2 discute as principais abordagens relacionadas com a tarefa de extração de aspectos em opiniões regulares e discute as principais contribuições científicas de cada abordagem. A Seção 3 descreve os experimentos realizados, incluindo os *Corpora* utilizados, as métricas usadas no processo de avaliação e os métodos implementados para cada abordagem. A Seção 4 apresenta os resultados dos experimentos e faz uma análise e discussão dos resultados. Finalmente, na Seção 5 são apresentadas as considerações finais juntamente com os trabalhos futuros.

2. Abordagens para Extração de Aspectos

A Figura 1 apresenta uma adaptação da taxonomia apresentada por Liu (2012) das principais abordagens usadas para realizar análise de sentimentos em nível de entidade-aspecto.

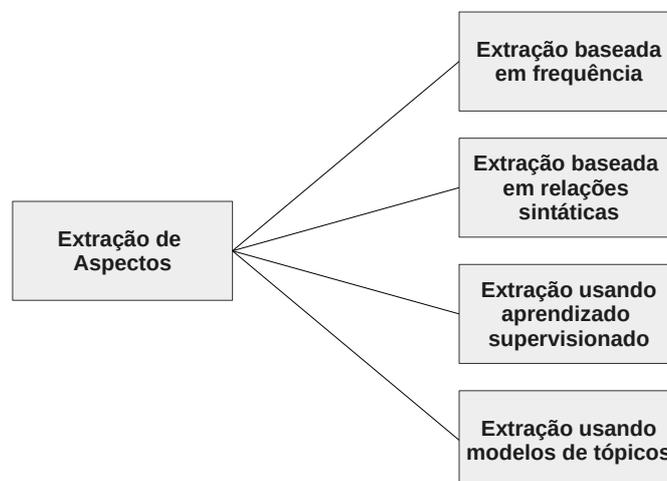


Figura 1. Taxonomia das principais abordagens usadas para extrair aspectos

2.1. Extração baseada em frequência

Esta abordagem visa encontrar expressões com aspectos explícitos representados por substantivos e frases nominais, a partir de um grande número de comentários em um dado domínio. Inicialmente, ela identifica os substantivos e frases nominais usando um etiquetador (*POS tagging*) e, em seguida, conta suas frequências e mantém apenas os mais frequentes, baseado em um limiar definido experimentalmente.

Segundo Hu & Liu (2004), a extração de aspectos baseada substantivos frequentes funciona porque os possíveis aspectos que são expressos como substantivos e frases nominais voltam a se repetir por meio de outros usuários que também estão relatando suas opiniões e experiências sobre os diferentes aspectos de uma entidade, os quais acabam convergindo para o mesmo vocabulário. Desse modo, substantivos e frases nominais que ocorrem com frequência são normalmente aspectos importantes e genuínos.

Uma recomendação é que o *Corpus* utilizado deve ter um número razoável de comentários do mesmo domínio, por exemplo, *smartphones*, filmes ou hotéis. Essa abordagem não funciona bem se o *Corpus* tiver uma mistura de produtos com grandes diferenças e/ou se cada produto tiver poucos comentários.

Scaffidi et al. (2007) propuseram um método para comparar a frequência de substantivos e frases nominais de um *Corpus* de comentários de produtos, com as taxas de ocorrências obtidas em um outro *Corpus* genérico para identificar aspectos verdadeiros. Aprimorando essa abordagem, Moghaddam & Ester (2010) estudaram a utilização de coocorrência de palavras, buscando identificar os aspectos mais frequentes e removendo os substantivos irrelevantes. A remoção baseou-se no número de ocorrências no *Corpus*, utilizando remoção de afixos e palavras comuns.

Long et al. (2010) extraíram os aspectos baseado na frequência e nas informações de distância. Primeiramente, eles encontraram o aspecto núcleo usando o método baseado em frequência e, depois, utilizaram as informações de distância de similaridade usadas em Cilibrasi & Vitányi (2007) para encontrar outras palavras relacionadas ao aspecto.

A extração baseada em substantivos frequentes é a forma mais simples de extrair aspectos e tem servido como *baseline* para métodos mais complexos. Por outro lado, a desvantagem dessa abordagem é que muitos erros de extração ocorrem devido a substantivos frequentes que não representam um aspecto real [Qiu et al. 2011].

Na tentativa de melhorar o desempenho da extração de aspectos, Li et al. (2015) propuseram um método de extração baseado em frequência com PMI-IR (do inglês, *pointwise mutual information–information retrieval*), que utiliza busca web para medir a semelhança semântica entre aspectos candidatos e entidades alvos. O mecanismo de pesquisa mantém rotineiramente a cobertura e a atualização de seu conteúdo, o que permite estimativas mais atualizadas e completas de semelhanças sem nenhum esforço extra.

Rana & Cheah (2018) propõem uma poda, em dois níveis, para eliminação de aspectos irrelevantes que afetam o desempenho das abordagens. A proposta consiste em: (i) calcular a frequência de cada palavra, selecionando os aspectos mais frequentes; e (ii) calcular a similaridade de palavras não frequentes, eliminando os aspectos que não estão semanticamente relacionados com a entidade em questão.

2.2. Extração baseada em relações sintáticas

Observando a relação entre o alvo e o sentimento, Hu & Liu (2004) utilizaram palavras de sentimento para identificar aspectos poucos frequentes. Esse relacionamento pode ser explorado para extrair entidades e aspectos, pois as palavras de sentimento são conhecidas por meio de recursos léxicos. Da mesma forma, se alguma palavra de sentimento for desconhecida, essa relação também pode ser usada para extrai-la.

Qiu et al. (2011) propuseram um método para resolver tanto a extração de aspecto quanto a extração de palavras de sentimento. Eles utilizaram uma gramática de dependência para descrever as relações sintáticas entre aspectos e palavras de sentimento. Além disso, eles consideraram que as relações ocorrem somente de forma direta e que aspectos são representados por substantivos e palavras de sentimento por adjetivos.

A abordagem de Sousa et al. (2015) utilizou a estrutura sintática das sentenças para identificar as características e seus respectivos qualificadores por meio de padrões linguísticos. Os autores consideraram os verbos como palavras opinativas, além dos adjetivos e advérbios. Eles utilizaram padrões linguísticos pré-definidos por Turney (2002) e algumas extensões para satisfazer o domínio de *smartphones*.

Baseando-se em árvores de dependência, Wang & Pan (2019) construíram uma RNN (do inglês, *Recursive Neural Network*), associando a estrutura sintática de cada sentença ao aprendizado de características para aprender as interações entre aspectos e palavras opinativas que são invariáveis ao domínio. O modelo posposto codifica as interações sintáticas, considerando que as palavras correspondentes tendem a ter funcionalidades semelhantes nos diferentes domínios. A RNN codifica cada relação de dependência em um vetor que atua como uma ponte para a transferência de conhecimentos entre diferentes domínios. Por fim, o recurso gerado pela RNN é passado como entrada para uma GRU (do inglês, *Gated Recurrent Unit*) para classificação de cada palavra.

No geral, a principal desvantagem desta abordagem é o fato de que alguns aspectos são expressões multipalavras e, muitas vezes, eles são implícitos. Entretanto, o método é bastante eficaz principalmente para identificar palavras de sentimento independentemente do domínio.

2.3. Extração usando aprendizado supervisionado

Essas técnicas tratam a extração de aspectos como uma tarefa de classificação e exigem grandes *Corpora* rotulados para treinamento [Liu 2012, Liu 2015]. Elas utilizam Modelo Escondido de Markov (do inglês, *Hidden Markov Model – HMM*) [Rabiner 1990], Campo Aleatório Condicional (do inglês, *Conditional Random Fields – CRF*) [Lafferty et al. 2001] e Rede Neural Convolutiva (do inglês, *Convolutional Neural Network – CNN*) [Poria et al. 2016].

Jin & Ho (2009) aplicaram um modelo léxico baseado em HMM para identificar padrões e extrair aspectos e expressões de opinião de relatos. Jakob & Gurevych (2010) aplicaram CRF em *Corpora* de diversos domínios. Como atributos, eles utilizaram *tokens*, *POS tagging*, dependência sintáticas, distância de similaridade e palavras de opinião.

Yu et al. (2011) usaram modelos baseados em Máquina de Vetor de Suporte (do inglês, *Support Vector Machine – SVM*) para classificar os aspectos em um relato. Cada instância representa um termo que é classificado como *aspecto* e *não-aspecto*. Todos os

aspectos classificados são extraídos e representados através de sumários. O método também identifica expressões e sinônimos de aspectos frequentes. Entretanto, o método apresenta uma limitação importante: a identificação de aspectos só ocorre a partir de uma lista limitada de substantivos frequentes.

Alternativamente, técnicas de aprendizado profundo têm mostrado desempenho promissor em muitas tarefas da PLN, incluindo análise de sentimentos [Collobert et al. 2011]. Poria et al. (2016) apresentaram a primeira abordagem de aprendizado profundo para a extração de aspectos usando uma CNN de sete camadas para classificar cada palavra de uma sentença opinativa em *aspecto* ou *não-aspecto*, resultando em uma melhoria significativa da acurácia sobre o estado da arte.

Em [Jeon et al. 2019], os autores introduziram um novo método para a extração de aspectos e classificação de sentimentos baseado em grafos, e alcançaram resultados superiores às *baselines* como Word2Vec, Doc2Vec e LDA (do inglês, *Latent Dirichlet Allocation*). No artigo, uma rede de coocorrência de palavras foi construída, a partir de um *Corpus*, a fim de computar uma pontuação de dispersão sobre palavra-aspecto e palavra-sentimento. Essa rede foi usada como entrada de uma rede neural *feedforward* para a classificação de aspectos e sentimentos.

2.4. Extração usando modelos de tópicos

Essas abordagens utilizam modelos estatísticos para identificar tópicos em grandes *Corpora*. Quando aplicadas ao domínio de produtos, esses tópicos podem corresponder aos diferentes aspectos, palavras opinativas, assuntos, especialidades ou locais citados no texto.

Os modelos de extração de tópicos probabilísticos baseiam-se na premissa de que os documentos consistem em uma mistura de tópicos, e um tópico é uma distribuição probabilística sobre as palavras. Dentre as técnicas encontradas na literatura para a extração de tópicos, as propostas baseadas em distribuição de *Dirichlet*, como LDA e pLSA (*probabilistic LDA*), se destacam pela qualidade dos resultados obtidos e pela boa interpretabilidade das dimensões extraídas.

O modelo pLSA, introduzido por Hofmann (2017), foi o primeiro a formalizar a extração de tópicos probabilísticos. Apesar de prover uma boa base para uma análise dos textos, ele apresenta dois problemas: i) o processo de geração dos tópicos para cada documento não é definido, o que exige uma determinada quantidade de parâmetros que cresce linearmente com a quantidade de documentos, e que pode levar ao *overfitting* dos parâmetros estimados; e ii) o modelo pLSA não determina uma forma natural de calcular as probabilidades relacionadas a um documento que não está no conjunto de treino [Blei et al. 2003, Kim et al. 2012].

Para evitar esses problemas, Blei et al. (2003) propuseram o modelo LDA como uma extensão do modelo pLSA, que propõe um modelo generativo probabilístico no qual os tópicos são definidos como uma distribuição de probabilidade sobre um vocabulário fixo de termos. Uma característica importante do LDA é que cada relato possui sua própria distribuição de tópicos e, assim, um mesmo relato pode estar relacionado com vários tópicos no qual cada tópico tem sua proporção de relevância. A distribuição dos tópicos em cada documento obedece à distribuição multivariada de *Dirichlet*. Esse modelo é uma das técnicas mais proeminentes para a extração de tópicos e se torna atrativo por descobrir grupos de termos que aparecem frequentemente juntos. Entretanto,

os tópicos podem conter aspectos e palavras opinativas, que normalmente são separados [Liu 2015].

Vários modelos de tópicos, que são extensões de LDA, têm sido propostos para realizar extração de aspectos. Branavan et al. (2009) propuseram um método que faz uso das descrições de aspectos encontradas nos campos prós e contras dos relatos. O modelo consiste em duas partes: a primeira agrupa as frases-chaves prós e contras em categorias de aspectos baseado na distribuição de similaridade; e a segunda cria um modelo de tópico que indica os aspectos no texto do relato.

Mukherjee & Liu (2012) utilizaram palavras-sementes definidas pelo próprio usuário sobre algum aspecto para produzir distribuições de tópicos. Os autores também apresentam uma estratégia para separar palavras opinativas e aspectos.

Um modelo LDA semântico de código misto independente da linguagem (lcms-LDA) foi proposto em Asnani & Pawar (2018) com o objetivo de melhorar os agrupamentos para a tarefa de extração de aspectos. Eles fizeram a associação semântica para distribuição de palavras utilizando o algoritmo LDA para recuperação de aspectos latentes coerentes independentes do idioma. A ideia do modelo é que o conhecimento dos recursos lexicais de sincronizações multilíngues adiciona similaridade semântica, enquanto as frequências de coocorrência de palavras nos modelos de tópicos compreendem a similaridade em nível sintático por meios estatísticos. Os experimentos mostraram uma melhora na coerência e distinção dos agrupamentos de aspectos comparado com outras técnicas de dados de código misto.

2.5. Considerações sobre os trabalhos relacionados

Nas subseções anteriores, apresentamos os principais trabalhos da literatura que exploram a análise de sentimentos em nível de entidade-aspecto, agrupando-os em quatro categorias.

Ressaltamos que o objetivo do nosso trabalho é fazer uma análise comparativa entre métodos das quatro abordagens para extrair aspectos em textos opinativos no domínio de produtos. A nossa intenção é avaliar os métodos implementados e analisar o comportamento das abordagens em Corpora de relatos nos idiomas português e inglês. Neste sentido, torna-se inviável fazer uma comparação com os trabalhos citados, pois eles possuem diversas perspectivas e utilizam *Corpora* de diferentes domínios.

Adicionalmente, recomendamos ao leitor o trabalho de Solanki et al. (2019) que faz uma comparação de três algoritmos de aprendizagem de máquina no contexto de análise de sentimentos: *naïve bayes*, *árvore de decisão* e *máquina de vetor de suporte*. Por fim, o trabalho de Zhang et al. (2018) fornece uma visão geral de aprendizado profundo (do inglês, *deep learning*) e apresenta uma pesquisa abrangente de suas aplicações atuais na análise de sentimentos.

3. Experimentos

Nos experimentos realizados, fez-se uso de dois *Corpora* de relatos sobre produtos: *SemEval* e *Buscapé*, que serão descritos a seguir. Além disso, discute-se também as métricas usadas no processo de avaliação e os métodos implementados para cada uma das abordagens.

3.1. *Corpus SemEval*

Este *Corpus* foi disponibilizado pelos organizadores do *Workshop SemEval-2014* para os desafios propostos na área de análise semântica computacional [Pontiki et al. 2014]. O evento propôs a realização de quatro subtarefas, a saber: i) identificação de aspectos; ii) identificação da polaridade dos aspectos; iii) identificação da categoria dos aspectos; e iv) identificação da polaridade da categoria dos aspectos.

É importante destacar que, neste trabalho, realizou-se experimentos considerando apenas a primeira sub tarefa proposta. Um exemplo de relato do *Corpus* pode ser visto na Listagem 1. Para a avaliação dos resultados, a equipe do SemEval-2014 dividiu o *Corpus* em duas partes, sendo aproximadamente 80% do *Corpus* usado para treinamento e 20% usado para testes. Assim, das 3.841 sentenças do *Corpus*, 3.041 foram usadas para treinar o modelo e 800 para testá-lo.

3.2. *Corpus Buscapé*

Para os experimentos com relatos escritos em língua portuguesa, foi necessário a criação de um *Corpus* anotado, devido à falta de recursos para esse idioma. O *Corpus* foi coletado no dia 15/11/2017, usando um *Web Crawler* e possui 26.062 *reviews*, correspondendo a todos os relatos encontrados no site Buscapé na categoria de *Celular e Smartphone*. A Figura 2 mostra um exemplo de avaliação fornecida pelo site Buscapé e a Figura 3 mostra um exemplo de relato de um consumidor. A estratégia utilizada para anotação foi seguinte:

1. Anotar manualmente os aspectos encontrados nos campos Prós e Contras das avaliações dos produtos;
2. Anotar automaticamente os relatos dos consumidores, com base nos aspectos extraídos na etapa anterior.

Os aspectos utilizados para anotar os relatos foram extraídos manualmente por três especialistas da área de PLN, a partir das avaliações coletadas sobre diferentes smartphones. Cada um dos especialistas recebeu 324 sentenças para anotar as palavras que eles consideravam aspectos. Vale ressaltar que uma sentença pode ter zero ou mais aspectos e que um aspecto pode reaparecer outras vezes nas avaliações.

Listagem 1. Exemplo de relato do *Corpus SemEval*

```
<sentence id="3359">
  <text>The pizza is the best if you like thin crusted
  pizza.</text>
  <aspectTerms>
    <aspectTerm term="pizza" polarity="positive"
      from="4" to="9"/>
    <aspectTerm term="thin crusted pizza"
      polarity="neutral" from="34" to="52"/>
  </aspectTerms>
  <aspectCategories>
    <aspectCategory category="food" polarity="positive"/>
  </aspectCategories>
</sentence>
```

Galaxy S6 Edge mostra porque é um dos smartphones mais rápidos e mais robustos do mercado

Mesmo com o lindo design inovador, as bordas não possuem funções tão práticas e a bateria ficou abaixo da média.

por Colaborador em 29/9/2015

Prós:

- Design inovador
- Excelente desempenho
- Tela acima da média
- Ótima câmera

Contras:

- Duração da bateria ficou abaixo do esperado
- Funções da tela curva são questionáveis

Figura 2. Avaliação fornecida pelo site Buscapé

Consumidor E-Bit deu a nota: ★★★★★ ✓ **Recomendo este produto** 22/04/2016

COMPRADOR VERIFICADO

Inovação total

Um novo mundo de possibilidades com o Galaxy S6 Edge! A bateria carrega muito rápido e possui boa durabilidade, o design é super inovador, o seu custo benefício é (com certeza) muito compensador, seu display é incrível e para finalizar a câmera é como cinematográfica!!!

#top

Achou esta opinião útil? ▲ 2 ▼ 0

Figura 3. Relato de um consumidor publicado no site Buscapé

As anotações das palavras com mais concordância entre os especialistas foram definidas como aspectos e, em caso de divergência entre os três, a decisão ficou a critério de um quarto especialista, considerado o juiz. Os três especialistas convergiram em 108 das 324 sentenças. Houve maioria simples em outras 92 sentenças e os três divergiram em outras 124 sentenças. Esse resultado comprova que a identificação de aspectos não se trata de uma tarefa trivial. Apesar disso, destaca-se que a maioria das divergências foi relacionada com a forma como o aspecto foi anotado, por exemplo, na sentença “*O que eu mais gostei foi da câmera traseira dupla*”, as três formas de identificação foram: *câmera*, *câmera traseira* e *câmera traseira dupla*. No entanto, após a atuação do especialista-juiz, obteve-se 84 expressões que representam aspectos e que foram usadas para etiquetar os relatos dos consumidores, criando o novo *Corpus*, denominado *Corpus Buscapé*.

Uma limitação dessa estratégia de anotação é a possibilidade de existirem aspectos que não foram citados nas avaliações fornecidas pelo site Buscapé, mas que foram mencionados nos relatos dos consumidores implicando em aspectos não anotados. Na Figura 3, por exemplo, o consumidor opina sobre os aspectos *durabilidade* e *custo benefício* que, por sua vez, não são mencionados na avaliação do Buscapé presente na Figura 2.

Para avaliação dos experimentos, resolveu-se adotar a mesma estratégia usada no *workshop SemEval-2014*, dividindo o *Corpus* em dois conjuntos para treinamento e testes. Porém, usou-se a proporção de 90% do *Corpus* para treinamento e 10% para testes, correspondendo a 23.456 e 2.606 relatos, respectivamente.

3.3. Métricas de avaliação

As métricas utilizadas para a avaliação dos métodos implementados são as mesmas usadas em aplicações tradicionais de classificadores, a saber:

1. **Acurácia (A)**: razão entre as classificações corretas, considerando os aspectos e não-aspectos, e o total de predições (soma de classificações corretas e incorretas);

$$A = \frac{VP + VN}{VP + VN + FP + FN}$$

2. **Precisão (P)**: razão entre os aspectos classificados corretamente e o total de aspectos preditos correta e incorretamente;

$$P = \frac{VP}{VP + FP}$$

3. **Revocação (R)**: razão entre os aspectos que foram corretamente classificados e o total geral de aspectos da amostra;

$$R = \frac{VP}{VP + FN}$$

4. **Medida-F (F)**: medida que busca o equilíbrio entre precisão e revocação, por meio de uma média harmônica entre elas.

$$F = 2 * \frac{P * R}{P + R}$$

onde VP, VN, FP, e FN referem-se a verdadeiro positivo, verdadeiro negativo, falso positivo, e falso negativo, respectivamente.

3.4. Métodos implementados

- 1) **Substantivos Frequentes (SF)**: implementou-se um algoritmo com os seguintes passos: carregar o *Corpus*; converter o texto para minúsculo; transformar o texto em *tokens* (*tokenização*); eliminar as contrações (por exemplo, “da” para “de a” ou “don’t” para “do not”); etiquetar cada um dos *tokens* com a respectiva classe gramatical (etiquetagem); extrair todos os substantivos e frases nominais usando expressões regulares; aplicar *stemmer* nas palavras extraídas; e, por fim, aplicar o algoritmo de mineração de dados Apriori [Agrawal & Srikant 1994] para a descobrir os aspectos frequentes por meio das regras de associações geradas. Usou-se a biblioteca NLTK³ (*Natural Language Toolkit*), os etiquetadores *Perceptron Tagger* para o *Corpus* em inglês e do *POS Tagger* para o *Corpus* em português, e os *stemmers Porter Stemmer* [Porter 1980] para o inglês e *RSLP Stemmer* [Orengo & Huyck 2001] para o português.
- 2) **Padrões Linguísticos (PL)**: O método implementado também fez uso da biblioteca NLTK para o procedimento de *tokenização* e etiquetagem, da seguinte forma: o texto foi convertido em uma sequência de *tokens*; depois, etiquetou-se os *tokens* usando um *POS Tagger*; por fim, extraiu-se as palavras cujas classes gramaticais correspondem aos padrões linguísticos. Para o *Corpus* em inglês, usou-se os padrões descritos por Turney (2002) (ver a Listagem 2) e para o *Corpus* em português, usou-se os padrões propostos por Sousa et al. (2015) (ver a Listagem 3).

³ Disponível em <http://www.nltk.org>

Listagem 2. Padrões linguísticos definidos por Turney (2002)

1. <ADJ><SUBS>
2. <ADV><ADJ>?!<SUBS>
3. <ADJ><ADJ>?!<SUBS>
4. <SUBS><ADJ>?!<SUBS>
5. <ADV><VERB>

Listagem 3. Padrões linguísticos definidos por Sousa et al. (2015)

1. <ADJ><SUBS> (<PREP>?<SUBS>)*
2. <ADV><ADV>?<ADJ> (<SUBS> (<PREP>?<SUBS>)*)*?
3. <SUBS> (<PREP>?<SUBS>)* (<ADJ>)<ADV><ADV>?
4. <SUBS> (<PREP>?<SUBS>)*<ADV>?<ADJ>+
5. <ADV><VERB>
6. <VERB><ADV>

3) Aprendizagem Supervisionada (CNN): o método implementado usa uma rede neural convolucional da biblioteca Keras⁴ e tem uma arquitetura básica com uma camada de entrada, uma camada de convolução, uma camada de *pooling*, *dropout* e uma camada totalmente conectada com saída *softmax*. A estrutura básica foi usada para evitar uma vantagem sobre os outros métodos implementados que também usam as configurações básicas. A Tabela 3 apresenta os parâmetros da CNN com base nos experimentos sobre parâmetros de Vieira & Moura (2017). A entrada da rede é uma matriz de *embeddings* com 300 dimensões. Em sua camada de convolução tem-se 100 filtros de comprimento 10. A camada de *pooling* segue a camada de convolução com *max-pooling* de tamanho 2. As saídas das camadas de convoluções foram computadas usando uma função de ativação.

A função de ativação *relu*, a função objetivo *categorical_hinge* e o algoritmo de otimização *adam* foram definidos após execução de uma série de experimentos com as opções disponíveis na biblioteca Keras.

Tabela 3. Valores dos parâmetros usados na CNN

Parâmetros	Valor
<i>embedding_dim</i>	300
<i>num_filters</i>	100
<i>filter_lenght</i>	10
<i>pool_size</i>	2
<i>batch_size</i>	8
<i>epochs</i>	30
<i>activation_function</i>	relu
<i>optimizer</i>	adam
<i>loss</i>	categorical_hinge

⁴ Disponível em <https://keras.io/>

- 4) **Modelo de tópicos (LDA):** implementou-se uma alocação latente de Directlet com algoritmo *Gibbs Sampling* para estimar os parâmetros do modelo e pequenas modificações entre inglês e o português, como a lista de *stopwords*, o *stemmer* e a remoção de contrações. Um experimento foi realizado para definir a quantidade de tópicos e observou-se que, quanto maior o número de tópicos, piores são os resultados. Por exemplo, com dois tópicos a precisão baixou em 1,34% e a revocação baixou em 12,64%, sempre utilizando um total de 50 palavras. Vale ressaltar que tópicos diferentes podem conter as mesmas palavras explicando a queda dos resultados. Os valores dos parâmetros usados nos demais experimentos foram: $\alpha=0.5$, $\beta=0.5$, número de tópicos=1 e iterações=50.

4. Resultados e Discussões

As Tabelas 4 e 5 mostram as métricas de avaliação calculadas para cada um dos algoritmos considerando os *Corpora* SemEval e Buscapé, respectivamente.

Observa-se a baixa precisão do método SF no *Corpus* Buscapé (25,41%) em consequência do grande número de falsos positivos. Após realizar uma análise dos erros nas palavras geradas pelo método, verificou-se que muitas delas não eram substantivos, ou seja, o etiquetador para o português gerou a classe gramatical incorreta. Isso certamente foi decorrente de erros ortográficos e/ou gírias usadas nos relatos. Assim, conclui-se que o método SF é totalmente dependente da forma como o *Corpus* está escrito e dos recursos computacionais utilizados, como etiquetadores e *stemming*. Já para o *Corpus* SemEval, obteve-se uma precisão de 42,15%. Com relação a revocação o método SF alcançou o melhor resultado do *Corpus* SemEval (63,60%) e o segundo melhor para o *Corpus* Buscapé (59,44%), comprovando que os substantivos frequentes são verdadeiros indicadores de aspectos. No entanto, esses resultados apontam para a importância de realizar pré-processamento dos dados textuais antes de usar ferramentas.

Como o método PL se baseia nas classes gramaticais das palavras para formar os padrões linguísticos, a identificação incorreta das *tags* pelos etiquetadores também prejudicou a precisão do método (ver as Tabelas 4 e 5), a saber: 50,42% para o *Corpus* SemEval e 26,09% para o *Corpus* Buscapé. Isso implica que esse método é dependente da qualidade de escrita do texto, como também foi comprovado por de Sousa et al. (2015).

Tabela 4. Comparação entre abordagens usando o *Corpus* SemEval.

Abordagem	Acurácia	Precisão	Revocação	Medida-F
SF	81.45%	42.25%	63.60%	50.77%
PL	85.02%	50.42%	21.52%	30.17%
CNN	89.50%	94.63%	31.98%	47.81%
LDA	79.69%	31.19%	29.08%	30.10%

Tabela 5. Comparação entre abordagens usando o *Corpus* Buscapé.

Abordagem	Acurácia	Precisão	Revocação	Medida-F
SF	89.21%	25.41%	59.44%	35.60%
PL	92.61%	26.09%	25.79%	25.94%
CNN	96.71%	96.76%	35.74%	52.20%
LDA	81.00%	16.59%	69.13%	26.76%

Após análise dos erros, alguns fatores podem justificar a incongruência entre os resultados: *i)* o etiquetador para a língua inglesa é melhor que o etiquetador da língua portuguesa; e *ii)* o *Corpus* Buscapé contém muitas palavras com erros ortográficos, prejudicando a etiquetagem. Em decorrência disso, o método PL obteve uma baixa revocação pois foram identificados poucos padrões linguísticos, fazendo o método “chutar” menos.

Com relação à abordagem supervisionada CNN, obteve-se resultados semelhantes nos dois *Corpora*: precisão de 94,63% no *Corpus* SemEval e 96,67% no *Corpus* Buscapé. No entanto, a revocação e a Medida-F foram baixas. Isso se deve porque as classes são bastantes desbalanceadas, apenas 15% das palavras do *Corpus* SemEval e 5% das palavras do *Corpus* Buscapé são *aspectos*. Todas as demais palavras são *não-aspectos*.

O modelo de tópicos LDA foi o que apresentou maior divergência entre os *Corpora*. Aplicado ao *Corpus* Buscapé, ele encontrou a maior revocação dentre os métodos analisados (69,13%), contudo, apresentou também a pior precisão (16,59%). Já no *Corpus* SemEval, ele obteve os piores resultados. Isso pode ser explicado pelo fato do *Corpus* SemEval possuir mais expressões que representam aspectos (1295) do que o *Corpus* Buscapé (84). Assim, os poucos verdadeiros positivos encontrados no Buscapé já alcançavam uma boa revocação, o que não vale para o SemEval. Além disso, a saída da LDA é formada por *unigramas* (palavras únicas) e as expressões que representam aspectos do *Corpus* SemEval possuem estruturas multipalavras. Adicionalmente, uma análise das palavras preditas como *aspectos* pelo modelo LDA foi feita e observou-se que as palavras selecionadas no modelo de tópicos incluíam palavras de sentimentos, entidades, e outros ruídos.

Analisando a acurácia que leva em consideração os *não-aspectos* corretamente classificados, diferentemente das outras medidas, destaca-se que o método de aprendizado supervisionado (CNN) foi o que menos errou, em contraste com o método LDA que teve o pior resultado nessa métrica.

Considerando a Medida-F, destaca-se que o método CNN obteve o maior valor para o *Corpus* Buscapé e que o método SF obteve o maior valor para o *Corpus* SemEval, seguido de perto do método CNN. Por fim, ressaltamos que nossos resultados são não conclusivos para apontar o melhor método para extração de aspectos em aplicações de análise de sentimentos baseada em aspectos. Porém, eles norteiam empresas e profissionais de TI sobre os problemas práticos encontrados em tais aplicações.

5. Considerações Finais e Trabalhos Futuros

Este artigo apresentou uma série de experimentos usados para avaliar e comparar os resultados de quatro abordagens utilizadas para extrair aspectos no problema de análise de sentimentos baseada em aspectos. Além disso, comparou-se também os quatro métodos implementados entre *Corpora* das línguas portuguesa e inglesa.

O método baseado em substantivos frequentes surpreendeu por ser um algoritmo simples e alcançar melhores resultados que as abordagens baseadas em modelos de tópicos (LDA) e relações sintáticas. A abordagem usando aprendizado supervisionado (CNN) alcançou resultados expressivos: precisão de 96,76% no *Corpus* Buscapé, mas com a revocação baixa, apenas 35,73% no mesmo *Corpus*. Certamente, isso aconteceu devido à falta de balanceamento das classes nas bases de relatos utilizadas.

Destaca-se também como uma contribuição deste trabalho, a criação de um *Corpus* com 26.062 comentários de opiniões de consumidores sobre produtos em português com anotação de aspectos, que está disponível em <[link removido para revisão às cegas](#)>. No entanto, como o *Corpus* foi criado de forma semiautomática, ele pode conter algumas limitações, tais como: aspectos não anotados, ausência de aspectos implícitos, opiniões comparativas, palavras com grafia incorreta, entre outros ruídos.

Como trabalhos futuros deseja-se: i) analisar os modelos discutidos em outros domínios, tais como: hotéis, filmes, e redes sociais; e ii) propor um modelo híbrido, combinando as técnicas SF e PL e o modelo CNN para tentar melhorar o desempenho.

Referências

- Agrawal, R. and Srikant, R. (1994) “Fast algorithms for mining association rules in large databases”. In: *VLDB’94, Proceedings of 20th International Conference on Very Large Data Bases*. 487-499. [[GS Search](#)]
- Asnani, K., and Pawar, J. D. (2018). Extraction of Code-mixed Aspect Topics in Semantic Representation. *Computación y Sistemas*, 22(1), 55-63. [[GS Search](#)]
- Barbosa, J., Moura, R. and Santos, R. (2016) “Predicting Portuguese Steam Review Helpfulness Using Artificial Neural Networks”. In: *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*. [[GS Search](#)]
- Bickart, B. and Schindler, R. (2001) “Internet forums as influential sources of consumer information”. *Journal of Interactive Marketing*, 15(3), 31-40. [[GS Search](#)]
- Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) “Latent dirichlet allocation”. *Journal of Machine Learning Research*, 3, 993-1022. [[GS Search](#)]
- Branavan, S., Chen, H., Eisenstein, J. and Barzilay, R. (2009) “Learning document-level semantic properties from free-text annotations”. In: *Journal of Artificial Intelligence Research*, 34, 569-603. [[GS Search](#)]
- Cambria, E. (2016) “Affective computing and sentiment analysis”. In: *IEEE Intelligent Systems*, 31(2), 102–107. [[GS Search](#)]
- Cambria, E., Poria, S., Gelbukh, A. and Thelwall, M. (2017). “Sentiment analysis is a big suitcase”. In: *IEEE Intelligent Systems*, 32(6), 74-80. [[GS Search](#)]
- Cilibrasi, R. and Vitányi, P. M. B. (2007) “The google similarity distance”. *IEEE Transactions on Knowledge and Data Engineering*, 19(3), 370-383. [[GS Search](#)]
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011) “Natural language processing (almost) from scratch”. *Journal of Machine Learning Research*, 12, 2493-2537. [[GS Search](#)]
- Cui, G., Lui, H. K., and Guo, X. (2014) “The effect of online consumer reviews on new product sales”. In: *Int. Journal of Electronic Commerce*, 17(1), 39-58. [[GS Search](#)]
- Fretwell, L., Stine, J., Sethi, H., and Noronha, A. (2013). “Catch and keep’digital shoppers: how to deliver retail their way”. In: *CISCO Internet Business Solutions Group*. [[GS Search](#)]
- Hofmann, T. (2017) “Probabilistic latent semantic indexing”. In: *ACM SIGIR Forum*, 51(2), 211-218. [[GS Search](#)]
- Hu, M. and Liu, B. (2004) “Mining and summarizing customer reviews”. In: *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 168-177. [[GS Search](#)]

- Jakob, N. and Gurevych, I. (2010) “Extracting opinion targets in a single and cross-domain setting with conditional random fields”. In: *Proc. of the Conference on Empirical Methods in Natural Language Processing*. 1035-1045. [[GS Search](#)]
- Jeon, S. W., Lee, H. J., Lee, H., and Cho, S. (2019). “Graph Based Aspect Extraction and Rating Classification of Customer Review Data”. In *Int. Conference on Database Systems for Advanced Applications*. Springer, Cham. 186-199. [[GS Search](#)]
- Jin, W. and Ho, H. H. (2009) “A novel lexicalized HMM-based learning framework for web opinion mining”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 465-472. [[GS Search](#)]
- Kim, H. D., Park, D. H., Lu, Y. and Zhai, C. (2012) “Enriching text representation with frequent pattern mining for probabilistic topic modeling”. *Proc. of the American Society for Information Science and Technology*, 49(1), 1-10. [[GS Search](#)]
- Lafferty, J. D., Mccallum, A. and Pereira, F. C. N. (2001) “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”. In: *Proc. of the 18th International Conference on Machine Learning*. 282-289. [[GS Search](#)]
- Li, S., Zhou, L., and Li, Y. (2015). “Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures”. *Information Processing & Management*, 51(1), 58-67. [[GS Search](#)]
- Liu, B. (2012) “Sentiment Analysis and Opinion Mining”. *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167. [[GS Search](#)]
- Liu, B. (2015) “Sentiment Analysis: mining sentiments, opinions, and emotions”. *Cambridge University Press*, 1 edition. [[GS Search](#)]
- Long, C., Zhang, J. and Zhu, X. (2010) “A review selection approach for accurate feature rating estimation”. In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. 766-774. [[GS Search](#)]
- Moghaddam, S. and Ester, M. (2010) “Opinion digger: an unsupervised opinion miner from unstructured product reviews”. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. 1825-1828. [[GS Search](#)]
- Mukherjee, A. and Liu, B. (2012) “Modeling review comments”. In: *Proc. of the 50th Annual Meeting of the Assoc. for Computational Linguistic*. 1, 320-329. [[GS Search](#)]
- Orengo, V. and Huyck, C. (2001) “A stemming algorithm for the portuguese language”. In: *Proc. 8th Symp. on String Processing and Information Retrieval*. [[GS Search](#)]
- Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S. (2014) “Semeval-2014 task 4: Aspect based sentiment analysis”. In *Proc. of the 8th International Workshop on Semantic Evaluation*. 27-35. [[GS Search](#)]
- Poria, S., Cambria, E. and Gelbukh, A. (2016) “Aspect extraction for opinion mining with a deep convolutional neural network”. *Knowledge-Based Systems*. [[GS Search](#)]
- Porter, M. F. (1980) “An algorithm for suffix stripping”. *Program*. [[GS Search](#)]
- Qiu, G., Liu, B., Bu, J. and Chen, C. (2011) “Opinion word expansion and target extraction through double propagation”. *Computational Linguistics*, 37(1), 9-27. [[GS Search](#)]
- Rabiner, L. (1990) “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. In: *Readings in speech recognition*. Morgan Kaufmann. 267-296. [[GS Search](#)]
- Rana T. and Cheah, YN. (2018) “Improving Aspect Extraction Using Aspect Frequency and Semantic Similarity-Based Approach for Aspect-Based Sentiment Analysis”. In: *Int. Conference on Computing and Information Technology*. 317-326 [[GS Search](#)]

- Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H. and Jin, C. (2007) “Red opal: product-feature scoring from reviews”. In: *Proceedings 8th ACM Conference on Electronic Commerce*. 182-191. [[GS Search](#)]
- Solanki, V. K., Nguyen H. H. C. and Zonghyu, J. L. (2019) “Opinion mining: using machine learning techniques”. In: *Extracting Knowledge From Opinion Mining. IGI Global*. 66-82 [[GS Search](#)]
- Sousa, R., Rabêlo, R. and Moura, R. (2015) “A fuzzy system-based approach to estimate the importance of online customer reviews”. In: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* 1-8. [[GS Search](#)]
- Turney, P. D. (2002) “Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews”. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 417-424. [[GS Search](#)]
- Vieira, J. P. A., and Moura, R. S. (2017) “An analysis of convolutional neural networks for sentence classification”. In: *XLIII Latin American Computer Conference (CLEI)* 1-5. [[GS Search](#)]
- Young, T., Hazarika, D., Poria, S. & Cambria, E. (2018). “Recent trends in deep learning based natural language processing”. In: *IEEE Computational intelligence magazine*, 13(3), 55-75. [[GS Search](#)]
- Yu, J., Zha, Z., Wang, M. and Chua, T. (2011) “Aspect ranking: Identifying important product aspects from online consumer reviews”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. 1496-1505. [[GS Search](#)]
- Wang, W. and Pan, S. (2019). “Syntactically-Meaningful and Transferable Recursive Neural Networks for Aspect and Opinion Extraction”. *Computational Linguistics*, 1-32. [[GS Search](#)]
- Zhang, L., and Liu, B. (2014) “Aspect and entity extraction for opinion mining”. In: *Data mining and knowledge discovery for big data* 1-40. [[GS Search](#)]
- Zhang, L., Shuai W. and Liu, B. (2018) “Deep learning for sentiment analysis: A survey”. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. [[GS Search](#)]