UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA
2012.2

UM MÓDULO DE INDEXAÇÃO DE CONTEÚDOS OPINATIVOS PARA O PAIRCLASSIF – UM SISTEMA DE CLASSIFICAÇÃO DE SENTIMENTO BASEADO EM PARES

Por

Wagner de Souza Rolim

Trabalho de Graduação

Recife, Maio de 2013

UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA
2012.2

UM MÓDULO DE INDEXAÇÃO DE CONTEÚDOS OPINATIVOS PARA O PAIRCLASSIF – UM SISTEMA DE CLASSIFICAÇÃO DE SENTIMENTO BASEADO EM PARES

Trabalho de Graduação apresentado ao Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do Grau de Bacharel em Engenharia da Computação

Orientadora: Profa. Dra. Flávia de Almeida Barros
Co-orientador: Nelson Gutemberg Rocha da Silva

Recife, Maio de 2013

Aos meus pais, Edvaldo e Eunice, e aos
meus irmãos, Walter, Wlademir e Walério

Agradecimentos

Primeiramente, agradeço a Deus, pois tudo que tenho e sou vem dEle. Por Seu amor, perdão e graça para comigo. Porque Ele é bom e Sua fidelidade não tem fim.

Aos meus pais, por terem me criado com muito amor, e ensinado princípios e valores que me transformaram em um homem de caráter. Em especial ao meu pai, Edvaldo de Souza Rolim, que infelizmente não viveu até ver um filho se formar em Engenharia.

Aos meus irmãos, cunhadas e sobrinhos, por proporcionarem um ambiente familiar tão unido.

À minha namorada, Carolina, pelo apoio, ânimo nos momentos de tristeza, e por me fazer feliz.

À minha orientadora, profa. Flávia Barros, que como uma mãe, nos momentos certos aconselhou, incentivou, deu bronca, e principalmente ajudou compartilhando seus conhecimentos.

Aos demais professores e funcionários do Centro de Informática, que contribuíram com minha formação acadêmica em um centro de referência.

Aos meus amigos e irmãos da IECP, que sempre estão prontos a me abraçar, interceder por mim, sorrir e chorar.

E finalmente, aos meus colegas de trabalho da Unisys e projeto e-Fisco, por proporcionarem um ambiente de trabalho tão prazeroso, em especial à minha equipe, pelo apoio nesta fase final da monografia.

A todos, muito obrigado!

"O temor do Senhor é o princípio da sabedoria"

Salomão

Resumo

Cada vez mais pessoas usam a internet para expressar, através de blogs, fóruns e redes sociais, as suas opiniões, sentimentos, avaliações acerca dos mais variados produtos, serviços, marcas, empresas e eventos.

Esse tipo de informação tanto é importante para as empresas, que podem aperfeiçoar seus produtos ou prestação de serviços diante do consumidor e da concorrência, como é importante para as pessoas terem um conhecimento prévio sobre algo de seu interesse que auxilie numa decisão de compra ou contratação. Porém, como essa informação está dispersa de forma não organizada por toda a Web, o processo de coletar, analisar e sumarizar um grande conjunto de opiniões é algo inviável.

A área da Análise de Sentimentos surgiu exatamente para automatizar esse processo, avaliando a subjetividade de um texto, extraindo características do objeto em análise, classificando a opinião e sumarizando os resultados.

Este trabalho tem por objetivo estender o método PairClassif de classificação baseada em pares, através da implementação de um módulo de indexação de conteúdos opinativos que será utilizado em uma fase intermediária do processo de classificação.

Palavras-chave: Análise de Sentimentos, Classificação de Opinião, PairClassif, Módulo de Indexação de Conteúdo.

Abstract

Each day, more and more people use the internet to express, through blogs, forums and social networks, their opinions, feelings, reviews on all kinds of products, services, brands, businesses and events.

This kind of information is important both for companies, which can improve its products or services in face of consumers and competition, as it is important for people who can have prior knowledge about something of their interest that assists decisions to purchase or contract. However, as this information is scattered in a non-organized mode throughout the Web, the process of collecting, analyzing and summarizing a large set of opinions is something impracticable.

The area of Sentiment Analysis came exactly to automate this process, evaluating the subjectivity of a text, extracting features of the object under analysis, classifying the opinion and summarizing the results.

This paper aims to extend the PairClassif method of peer-based classification, through the implementation of an opinionated content indexing module that will be used in an intermediate stage of the classification process.

Keywords: Sentiment Analysis, Opinion Classification, PairClassif, Content Indexing Module.

Sumário

1. Introdução

Neste capítulo, serão descritos o contexto e a motivação que conduziram a preparação deste trabalho, bem como o que se espera obter ao final da implementação, além da estrutura deste documento.

1.1 Contexto e Motivação

Em 2012, o número de internautas ao redor do mundo atingiu a marca de 2,4 bilhões, havendo um expressivo consumo de mídias sociais e disseminação de informação [1][2]. Os usuários compartilham experiências, críticas e opiniões sobre os mais variados temas, inclusive produtos e serviços.

A partir dessa gama de opiniões, as empresas podem extrair dados que diagnostiquem o quão bem sucedidos os seus produtos estão no mercado. Por outro lado, os relatos de outros usuários sobre determinado produto, marca ou serviço são muito válidos para se levar em consideração quanto a adquirir ou não o que se deseja.

Mas devido à desorganização destas informações dispersas na Web, o processo de encontrar, analisar e classificar um conjunto de opiniões torna-se uma tarefa bem complexa. Nem os engenhos de busca atuais são capazes de fornecer uma síntese com as opiniões sobre um determinado item, ou mesmo relatos referentes à comparação entre itens.

Neste cenário, surgiu a área de pesquisa denominada Análise de Sentimentos [9], cujo objetivo principal é permitir que um usuário obtenha um relatório contendo o que as pessoas andam dizendo sobre algum item sem precisar encontrar e ler todas as opiniões e notícias a respeito, classificando os textos opinativos (textos que contém opiniões sobre o objeto ou tópico de interesse) com uma polaridade positiva, negativa ou neutra.

As atividades concernentes à Análise de Sentimentos dividem-se em:

- **Coleta**: visa buscar na web conteúdos sobre o item de interesse, e também identificar se esse conteúdo é relativo a um fato ou uma opinião. Fatos devem ser descartados, já que o interesse é nas opiniões dos usuários.
- **Extração**: determina os objetos descritos na opinião e as respectivas características comentadas.
- **Classificação**: atribui uma polaridade às opiniões recuperadas, classificando-as como positivas, negativas ou neutras.
- **Sumarização**: as classificações das diversas opiniões devem ser sumarizadas para o usuário, com o intuito de facilitar o seu entendimento sobre as mesmas. Esta sumarização pode ser em forma de texto ou gráfico.

Neste trabalho, será abordada a etapa referente à classificação de sentimentos.

1.2 Objetivos

Este trabalho de graduação foi desenvolvido no escopo de uma pesquisa na área de Análise de Sentimento cujo objetivo é classificar opiniões em nível de característica. Essa pesquisa já teve como resultado inicial o desenvolvimento do sistema SAPair [3]. Esse sistema conta com dois módulos principais: um módulo de extração de característica e um módulo de classificação de sentimentos (o PairClassif).

Nesse contexto, este Trabalho de Graduação teve por objetivo melhorar o desempenho do PairClassif, apresentado na dissertação de Mestrado de Nelson Silva [6]. Para tanto, foi implementado um módulo de indexação de conteúdo de

textos opinativos, que é usado na etapa de Filtro de Polaridade, para substituir as consultas a um engenho de busca feitas para validar a classificação da polaridade da opinião sob análise. Mais detalhes sobre o SAPair e o PairClassif serão vistos no capítulo 2.

1.3 Estrutura do Trabalho

Além deste capítulo introdutório, este trabalho está dividido da seguinte maneira:

O capítulo 2 apresentará conceitos da área de Análise de Sentimentos, as etapas que a constituem, com ênfase na etapa de classificação, através da descrição resumida do método PairClassif.

O capítulo 3 descreverá o processo de implementação do módulo indexador de conteúdo a ser utilizado neste método, no qual foi utilizada a biblioteca Lucene, para recuperação de informação. Em seguida, será discutida a avaliação de desempenho do módulo indexador em comparação com o PairClassif.

E finalmente, o capítulo 4 trará uma conclusão deste trabalho, incluindo as dificuldades encontradas e trabalhos futuros.

2. Análise de Sentimentos

A Análise de Sentimentos (AS), também conhecida como Mineração de Opinião, é o campo de estudo que analisa as opiniões das pessoas, sentimentos, avaliações, atitudes e emoções relativas a entidades, tais como produtos, serviços, organizações, pessoas, problemas, eventos, tópicos e seus atributos [8], visando uma classificação (positiva, negativa ou neutra) destes. Para isso, são empregados métodos e técnicas de outras áreas como Processamento de Linguagem Natural, Recuperação de Informação, e Inteligência Artificial.

Uma análise manual já acontece por parte de usuários que buscam na Web comentários de pessoas que tiveram experiência com objetos do interesse. Mas desse modo a quantidade de informação consultada é muito limitada, e pode não refletir uma opinião global, o que identifica um nicho para exploração dos sistemas de AS. Outro nicho são as empresas, que podem saber a reputação de seus produtos e serviços, para melhorá-los e assim obter lucros.

Neste capítulo, será feita uma breve apresentação dos conceitos básicos de AS, as etapas que a constituem, e será feita uma breve descrição do método PairClassif, no qual este trabalho se baseia.

2.1 Conceitos Básicos

Para facilitar o entendimento deste trabalho, serão apresentados alguns conceitos da área de Análise de Sentimentos. O exemplo a seguir servirá como guia na compreensão destes conceitos.

> *"(1) Fiz uma compra na loja online da Walmart. (2) (2.1) Ela é uma loja confiável e (2.2) lá regularmente aparecem promoções muito boas. (3) (3.1) A entrega foi incrivelmente rápida, mas (3.2) a embalagem chegou amassada."*

Exemplo 2.1 Comentário criado pelo autor

Neste exemplo, nem todas as sentenças são classificáveis, pois nem todas detêm opinião do autor acerca do objeto sob análise. Essa primeira verificação traz os conceitos de:

- **Objeto**: qualquer entidade (produto, serviço, pessoa, empresa, evento) que é analisada pelo autor da opinião. No exemplo 2.1, o objeto analisado é a loja Walmart.
- **Sentença Objetiva**: texto que contém um fato ou informação sem o ponto de vista do autor. A sentença (1) do exemplo 2.1 representa uma sentença objetiva.
- **Sentença Subjetiva**: também nomeado texto opinativo, expressa a visão do autor sobre um objeto. As sentenças (2) e (3) do exemplo 2.1 se encaixam nesta definição.

Sentenças objetivas, tal qual a sentença (1), não são processadas em AS. As demais sentenças carregam opinião sobre um objeto, e suas respectivas características. A seguir seguem os demais conceitos:

- **Opinião**: é um ponto de vista, atitude, sentimento, emoção ou avaliação acerca de um objeto ou suas características, expressas por um autor, e que podem ter caráter positivo, negativo ou neutro.
- **Polaridade**: também é chamada de orientação de opinião, representa o caráter positivo (+1), negativo (-1) ou neutro (0) da avaliação feita pelo autor.
- **Característica**: também conhecida como aspecto, é um atributo, propriedade, parte ou componente de um objeto. No exemplo 2.1, existem as características "reputação" (implícita na sentença 2.1), "promoções", "entrega" e "embalagem" (explícitas).
- **Palavra opinativa**: são termos que qualificam as características. Em sua maioria são adjetivos e advérbios, e podem ter conotação positiva ou negativa em relação a característica.

Finalmente, é preciso distinguir os níveis de classificação em Análise de Sentimentos [7].

- **Classificação em Nível de Documento**: considera-se a opinião no escopo de todo documento ou sentença em questão. No exemplo 2.1, apenas a opinião na sentença (3.2) foi negativa, enquanto as demais foram positivas, logo o texto será classificado como uma opinião positiva.
- **Classificação em Nível de Característica**: cada característica é analisada isoladamente, o que garante uma classificação mais refinada do texto. No exemplo 2.1, as características "reputação", "promoções" e "entrega" tiveram avaliação positiva, mas a característica "embalagem" teve avaliação negativa.

Conhecidos os conceitos básicos de AS que serão utilizados neste trabalho, é importante explicar que tarefas fazem parte do processo de Análise de Sentimentos.

2.2 Etapas da Análise de Sentimentos

Um sistema completo de Análise de Sentimentos é dividido em quatro etapas, que normalmente são sequenciais e complementares. Esta divisão de tarefas visa a obtenção de melhores resultados de avaliação. As etapas são: Análise de Subjetividade, Extração de Características, Classificação de Sentimentos e Visualização e Sumarização.

- **Análise de Subjetividade**: é necessário identificar se uma sentença é subjetiva ou objetiva. Apenas as sentenças subjetivas devem ser consideradas, por carregarem opiniões passíveis de análise. Sentenças objetivas devem ser descartadas pelo sistema, pois fazer uma análise destas pode interferir na qualidade final do processo.

- **Extração de Características**: uma vez identificadas as sentenças subjetivas, é importante quais características do objeto estão sendo avaliadas pelo autor da opinião. A classificação de um objeto dependerá da polaridade presente no maior número de características deste objeto referenciadas na sentença. Esta é uma das tarefas mais difíceis da AS, menos propensa à automação.
- **Classificação de Sentimentos**: após a extração das características do objeto presentes na opinião, o próximo passo é fazer a classificação do sentimento, ou seja, a atribuição das polaridades às palavras opinativas. Esta pode ser feita em diferentes níveis de granularidade (característica, sentença e documento), sendo a classificação em nível de característica a mais refinada, e a única na qual é necessário haver a extração de características.
- **Visualização e Sumarização**: a última etapa do processo de Análise de Sentimentos consiste em agregar e representar os resultados numa exibição simples e clara para o entendimento do usuário. Ela pode ser feita basicamente de duas formas: textual, mais direcionada para análise de um único objeto, em que acontece uma sumarização baseada na extração de textos similares sobre cada característica e seus pontos positivos e negativos; ou gráfica, ideal para comparações entre produtos do mesmo domínio, sendo exibidos o quão positivas e negativas cada característica é.

2.3 O método PairClassif

PairClassif é um método para classificação de sentimento baseado em Lingüística e Estatística que foi implementado por Nelson Gutemberg Rocha da Silva e que integra o sistema SAPair, um sistema completo de Análise de

Sentimentos em nível de característica. Este método classifica a polaridade de pares no formato (característica, palavra opinativa), tendo sido proposto para preencher uma lacuna na etapa de classificação de sentimentos, pois até então uma palavra opinativa possuía apenas uma polaridade, independente da característica a qual se referisse, e foi apresentado em janeiro de 2013.

O SAPair é composto por três módulos principais: Extração de Pares [5], Classificação de Sentimento e Módulo de Visualização. Sua arquitetura é mostrada na figura 2.1 a seguir.

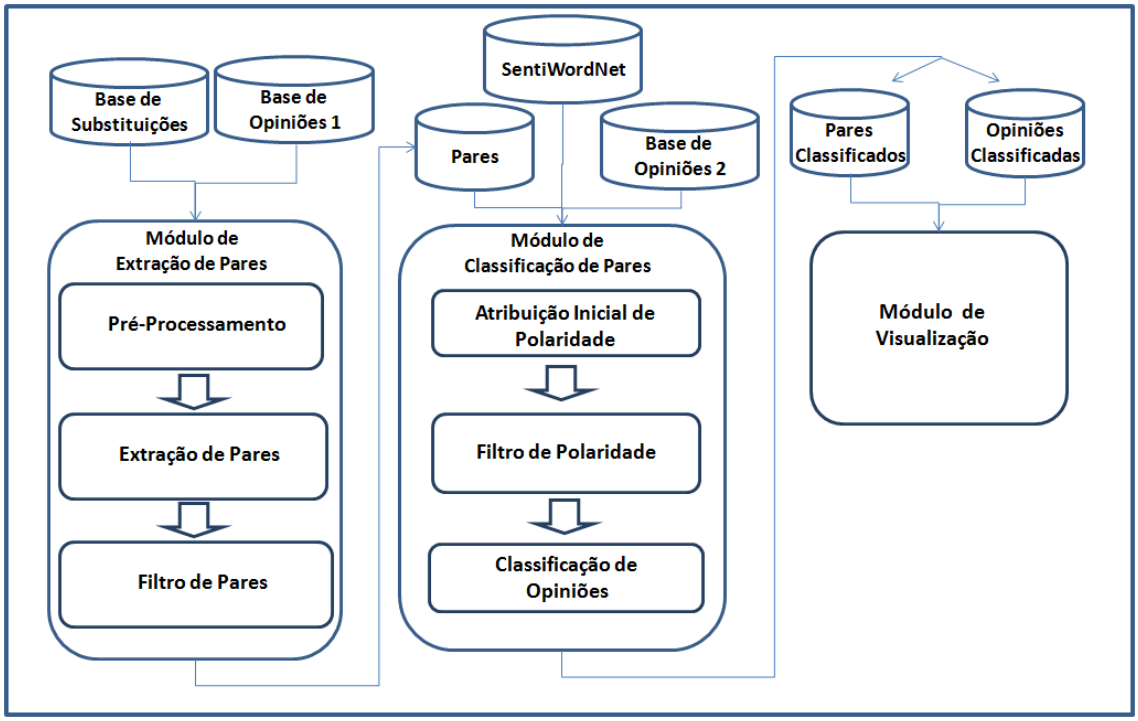

Figura 2.1 – Arquitetura do SAPair

Como este trabalho se baseia no módulo de classificação, não serão detalhados os demais módulos.

A metodologia PairClassif, para classificação de sentimento baseada em pares, é implementada em três etapas, a saber: Atribuição Inicial de Polaridade, Filtro de Polaridade, e Classificação de Opinião. Estas etapas estão organizadas de acordo com arquitetura mostrada na figura 2.2.

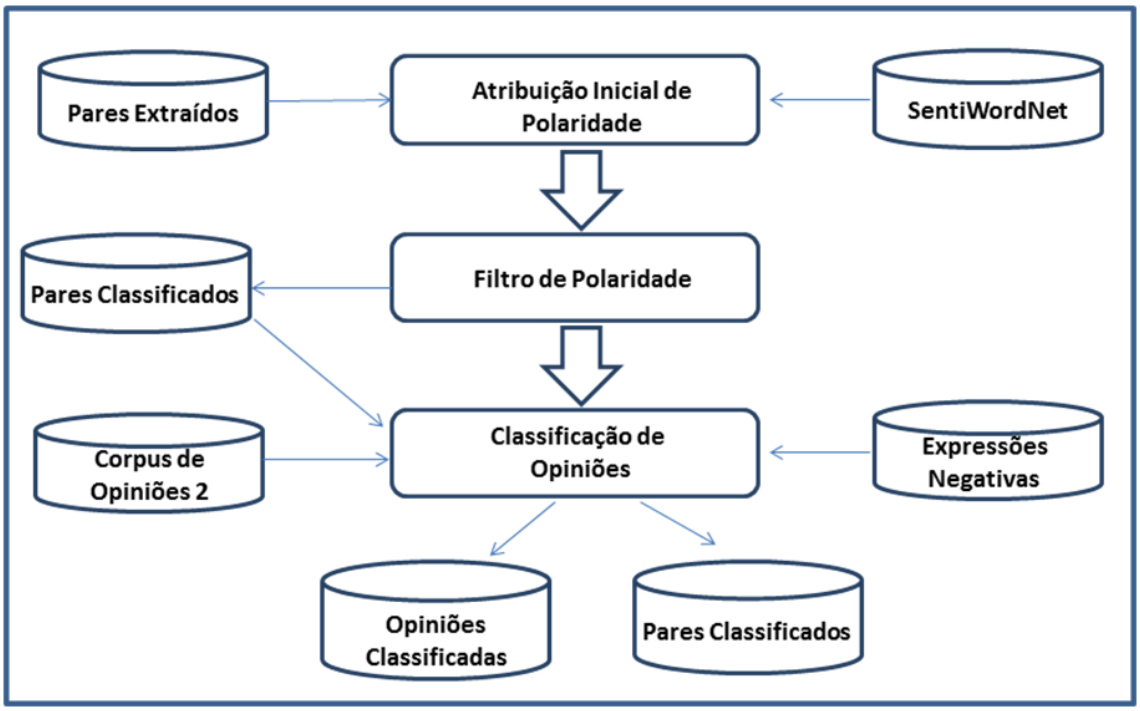

Figura 2.2 – Arquitetura do PairClassif

Na etapa de Atribuição Inicial de Polaridade, a palavra opinativa que compõe o par é polarizada com base no SentiWordNet (SWN) [11], uma ferramenta linguística disponível na Web e bastante usada no meio acadêmico, que possui uma extensa base de palavras com polaridade atribuída, havendo variação de scores positivos e negativos de acordo com o cenário. No PairClassif, para cada palavra opinativa é calculada a média de scores positivos e a média de scores negativos. Destas duas, a maior média é definida como a polaridade da palavra, e consequentemente do par. Se os scores forem iguais, a polaridade é neutra.

Na etapa de Filtro de Polaridade, acontece a validação da polaridade definida no passo anterior. Essa validação é feita porque existem palavras que têm score geral negativo, mas dependendo do cenário ela deveria ter score positivo (ou vice-versa). Por exemplo, de acordo com o SWN, a palavra "cold" apresenta uma polaridade geral negativa. Para o par "cold pizza", essa polaridade estaria correta,

porém, no cenário "the beer is cold" a palavra opinativa "cold" possui uma conotação positiva.

A validação ocorre pelo processo de autoria de Silva [6] denominado Classificação baseada em padrões, no qual são formados padrões de sentenças positivos e negativos para cada par, e estes padrões são consultados em engenho de busca na Web. O formato dos padrões encontra-se na seção 4.4.2 da dissertação.

A partir da quantidade de links retornados nas consultas, a polaridade do par é determinada através da fórmula $P = (P_0 - N_0) / (P_0 + N_0)$, onde P_0 é a quantidade de links retornados para os padrões positivos e N_0 a quantidade de links retornados para os padrões positivos. O valor da polaridade P varia entre -1 e 1.

A última etapa, Classificação de Opiniões, recebe como entrada uma base de opiniões do mesmo domínio da base utilizada no módulo de extração de pares e os pares classificados na etapa do Filtro de Polaridade. Ela compreende passos de Identificação dos pares na base de opiniões, Tratamento de Expressões Negativas, e Classificação da Opinião.

Primeiramente, essa base de opiniões é pre-processada, substituindo contrações, abreviações e gírias, e subdividindo as sentenças, de acordo com a pontuação (ponto, vírgula, ponto e vírgula, interrogação e exclamação). Depois disto, os pares são pesquisados nas sentenças.

Identificados os pares e suas polaridades, acontece o tratamento de expressões negativas. Quando a expressão relaciona-se ao par, a polaridade é invertida.

Por último, com as polaridades dos pares corrigidas/validadas, a classificação da opinião, podendo ser em nível de característica (foco da dissertação) ou em nível de documento. No nível de característica, para cada uma identificada numa opinião, contabiliza-se quantas vezes que ela foi classificada positivamente (apareceu em par de polaridade positiva), negativamente ou como neutra. Ao final do processo, sabe-se o número de vezes que tal característica

obteve determinada polaridade. E este número serve para a análise da classificação em nível de característica, feita no módulo de visualização do SAPair.

3. Módulo de indexação de conteúdos opinativos para o PairClassif

Como foi visto no capítulo anterior, o método PairClassif tem uma etapa chamada Filtro de Polaridade, na qual as polaridades inicialmente atribuídas através do SentiWordNet são validadas através de consultas de padrões de sentenças feitas a um engenho de busca (classificação baseada em padrões).

Como melhoramento ao sistema SAPair, este trabalho propõe restringir a consulta de validação da polaridade dos pares a um base de opiniões de domínio específico, ao invés de consultar a Web via um engenho de busca. Essa consulta mais restrita objetiva melhorar os resultados de classificação, uma vez que a precisão das consultas à Web inteira não é tão boa.

3.1 Base de Opiniões de Domínio Específico

Este TG utilizou duas bases de opiniões coletadas por Nelson Silva contendo milhares de opiniões, sendo uma no domínio de aparelhos celulares, e outra no domínio de filmes. Essas opiniões foram inicialmente armazenadas em arquivos Excel.

A partir desses arquivos, foi utilizada a biblioteca Apache POI [12], para converter cada opinião em um arquivo de texto único. Para cada avaliação (um usuário), a planilha Excel contém um campo de comentário livre, um campo de comentário favorável (Prós) e um campo de comentário desfavorável (Contras).

Se houver comentário livre, este é indexado. Caso contrário, é verificado se há comentário favorável ou desfavorável. Se sim, cada um é indexado. Portanto cada linha da planilha (correspondendo a uma avaliação) pode gerar um ou dois arquivos de texto no índice.

Para o domínio de aparelhos celulares, foram gerados 7149 arquivos, enquanto que para o domínio de filmes foram gerados 19173 arquivos.

3.2 Padrões de sentenças

Outro arquivo Excel, contendo características, respectivas palavras opinativas e os pares resultantes, foi utilizado para a formação dos padrões de sentenças. Para o domínio de aparelhos celulares, 216 pares foram formados, enquanto que para o domínio de filmes 209 pares foram formados.

Os pares são convertidos do Excel para arquivo de texto, e cada par é acrescentado de termos claramente positivos ou negativos, assim constituindo os padrões de sentenças a serem utilizadas como consultas (queries) ao índice. Os padrões têm o seguinte formato, usando o exemplo do par "cheap battery":

1. Positivos
   (i) cheap battery good amazing wonderful nice incredible wondrous
   (ii) cheap battery not bad poor wrong nasty unpleasant terrible awful dreadful under
2. Negativos
   (i) cheap battery bad poor wrong nasty unpleasant terrible awful dreadful under
   (ii) cheap battery not good amazing wonderful nice incredible

Esses padrões foram inspirados nos padrões utilizados por Nelson Silva em sua dissertação para consultar o engenho Google. Porém, o formato de apresentação das consultas foi modificado, para se adequar à sintaxe de consultas da biblioteca utilizada neste trabalho.

3.3 Construção do Índice e busca textual com o Lucene

Para poder indexar os milhares de arquivos de opiniões gerados e realizar as buscas dos padrões de sentenças no índice, foi utilizada a biblioteca Lucene [13].

O Apache Lucene, ou simplesmente Lucene, é um projeto mantido pela Apache Software Foundation que desenvolve software de busca open source e provê uma API (*Application Programming Interface*) com recursos para indexação, consulta e visualização de documentos, sendo licenciado através da Apache Software License. Inicialmente desenvolvido na linguagem Java, atualmente existem versões do Lucene para diversas outras linguagens, tais como C, C++, Delphi, .NET, Perl, PHP, Python e Ruby. Neste trabalho foi utilizada a versão 3.6.2 do Lucene em Java.

A construção do índice de documentos se deu da seguinte forma: cada documento do índice é composto por um nome e uma opinião, onde apenas a opinião é indexada, e consequentemente pode ser consultada. Já o nome é representado por um número que serve para identificar a opinião dentro da base de dados, não podendo ser consultado. Para permitir que as consultas aos padrões de sentenças com a presença do termo "not" tivessem êxito, foi usado uma instância de analisador que não retirava stopwords.

Uma vez que todos os documentos contendo opinião são adicionados ao índice, procede-se à consulta. Foram realizados dois experimentos. No primeiro, a consulta foi implementada de forma que se procura pelo par e algum dos termos subsequentes. Nas consultas onde ocorre o termo negativo "not", a pesquisa se dá igualmente pelo par e algum dos termos subsequentes, porém precedidos pelo "not". Numa notação lógica, a representação é:

- Consulta sem termo negativo: par AND (termo1 OR termo2 OR ..... OR termoN)
- Consulta com termo negativo: par AND (not termo1 OR not termo2 OR ..... OR not termoN)

No segundo experimento, a pesquisa ocorre restringindo a distância entre o par e cada um dos termos subsequentes, visando garantir que, por estarem fisicamente próximos, o termo positivo ou negativo realmente está se referindo ao par. Identicamente ao primeiro experimento, quando o padrão de sentença tem termo negativo, o formato da consulta apenas sofre o acréscimo deste termo. A representação é a seguinte:

- Consulta sem termo negativo: “par termo1”~1 OR “par termo2”~1 OR ..... OR “par termoN”~1
- Consulta com termo negativo: “par not termo1”~1 OR “par not termo2”~1 OR ..... OR “par not termoN”~1

Onde a notação ~X significa que os termos entre aspas tem no máximo X termos entre eles. Lembrando que o par tem a forma (característica, palavra opinativa).

Nos dois experimentos, foram utilizados objetos das classes BooleanQuery e PhraseQuery [14], do Lucene. Com a primeira classe, foi possível unir as várias cláusulas que existiam nos padrões numa mesma consulta, usando lógica booleana. Com a segunda classe, foi possível unir os termos que compunham cada par, pois eles deveriam obrigatoriamente aparecer juntos nos resultados das consultas.

Após a execução dos dois experimentos, percebeu-se que as consultas usadas no primeiro experimento retornaram mais registros do que as consultas do segundo experimento. Porém, em ambos os casos, a incidência de registros nas consultas com termos negativos foi menor que 10% da quantidade de pares.

Para se visualizar os resultados das consultas dos padrões de sentenças, é gerado um arquivo de texto que contém as informações do par pesquisado, cada padrão de sentença positivo e negativo deste par e os registros encontrados para cada consulta. De posse deste arquivo, é feita a contabilização tanto da quantidade

de registros concordantes com a polaridade, que foi determinada manualmente, como da quantidade de registros discordantes.

Vale ressaltar que a avaliação de desempenho do módulo indexador, que virá em seguida, foi calculada em cima dos resultados do primeiro teste.

3.4 Avaliação de desempenho

Uma vez contabilizados os registros (links) retornados tanto pelas consultas aos padrões positivos como pelas consultas aos padrões negativos, calcula-se então a polaridade P de cada par, cuja fórmula encontra-se no início deste capítulo.

Para avaliar o desempenho do módulo de indexação de conteúdo perante as consultas ao engenho de busca utilizadas na metodologia do PairClassif, foram utilizadas as mesmas métricas tradicionais para Classificação de Texto usadas na dissertação, a saber: Precisão (Precision), Cobertura (Recall) e F-measure [10]. No âmbito da Recuperação de Informação, a Precisão é a fração de documentos retornados que são relevantes para a consulta, enquanto que a Cobertura é a fração dos documentos que são relevantes para a consulta e que são retornados com sucesso. E a medida F-measure consiste na média harmônica entre a Precisão e a Cobertura. Estas foram calculadas para cada classe de polaridade (positiva, negativa ou neutra). A tabela 3.1 a seguir exemplifica, no formato de matriz confusão, os casos de acertos e erros possíveis nesta classificação.

		Automático		
		Classe 1 (+)	Classe 2 (-)	Classe 3 (0)
Manual	Classe 1 (+)	$t_p(1)$	erro(2,1)	erro(3,1)
	Classe 2 (-)	erro(1,2)	$t_p(2)$	erro(3,2)
	Classe 3 (0)	erro(1,3)	erro(2,3)	$t_p(3)$

Tabela 3.1 Modelo de matriz confusão para classificação em nível de característica

Onde $t_p(i)$ (true positive) é a quantidade de pares da classe i classificados corretamente, e erro(j,i) é a quantidade de pares da classe i classificados erroneamente como sendo da classe j.

Desta feita, as medidas de Precisão, Cobertura e F-measure são calculadas através das seguintes fórmulas:

Precisão = $t_p(i) / [t_p(i) + erro(i,j) + erro(i,k)]$

Cobertura = $t_p(i) / [t_p(i) + erro(j,i) + erro(k,i)]$

F-measure = 2*Precisão*Cobertura / (Precisão + Cobertura)

Agora, preenchendo a matriz confusão com os erros e acertos na classificação em nível de característica para os pares em cada um dos domínios testados com o módulo indexador, e fazendo a comparação com a classificação realizada através do método PairClassif, do sistema SAPair, tem-se:

(i) Para o domínio de aparelhos celulares

Módulo Indexador				
		Automático		
		Classe 1 (+)	Classe 2 (-)	Classe 3 (0)
Manual	Classe 1 (+)	80	4	76
	Classe 2 (-)	6	14	36
	Classe 3 (0)	0	0	0

Tabela 3.2 Matriz confusão da classificação em nível de característica usando o Módulo Indexador no domínio de aparelhos celulares

SAPair				
		Automático		
		Classe 1 (+)	Classe 2 (-)	Classe 3 (0)
Manual	Classe 1 (+)	155	5	1
	Classe 2 (-)	13	35	1
	Classe 3 (0)	0	0	0

Tabela 3.3 Matriz confusão da classificação em nível de característica usando o SAPair no domínio de aparelhos celulares

(ii) Para o domínio de filmes

Módulo Indexador				
		Automático		
		Classe 1 (+)	Classe 2 (-)	Classe 3 (0)
Manual	Classe 1 (+)	55	3	95
	Classe 2 (-)	9	12	35
	Classe 3 (0)	0	0	0

Tabela 3.4 Matriz confusão da classificação em nível de característica usando o Módulo Indexador no domínio de filmes

SAPair				
		Automático		
		Classe 1 (+)	**Classe 2 (-)**	**Classe 3 (0)**
Manual	Classe 1 (+)	128	18	0
	Classe 2 (-)	30	25	0
	Classe 3 (0)	0	0	0

Tabela 3.5 Matriz confusão da classificação em nível de característica usando o SAPair no domínio de filmes

Analisando as tabelas, percebe-se que as linhas referentes à classe 3 têm valores sempre iguais a 0 pois nenhum comentário foi manualmente classificado como neutro. Já na classificação automática, através das consultas ao módulo indexador, um número equivalente a 51,85% (112) dos 216 pares do domínio de aparelhos celulares, e 62,20% (130) dos 209 pares do domínio de filmes foi classificado como sendo neutro. Uma explicação para esta alta taxa é que pelo fato do indexador ser de tamanho muito inferior ao engenho de busca Google, utilizado no SAPair, muitos pares não foram encontrados juntos fisicamente nos documentos indexados. Outra explicação para esse resultado pode ser a formatação das consultas (padrões), que não correspondem às sentenças de fato postadas pelos usuários.

As tabelas a seguir mostram os valores calculados das medidas de Precisão, Cobertura e F-measure na classificação utilizando o módulo indexador e o sistema SAPair.

(i) Para o domínio de aparelhos celulares

Módulo Indexador			
	Classe 1 (+)	Classe 2 (-)	Classe 3 (0)
Precisão	93,02%	77,78%	0,00%
Cobertura	50,00%	25,00%	0,00%
F-Measure	65,04%	37,84%	0,00%

Tabela 3.6 Desempenho da classificação em nível de característica usando o Módulo Indexador no domínio de aparelhos celulares

SAPair			
	Classe 1 (+)	Classe 2 (-)	Classe 3 (0)
Precisão	92,3%	87,5%	0,0%
Cobertura	96,3%	71,4%	0,0%
F-Measure	94,2%	78,7%	0,0%

Tabela 3.7 Desempenho da classificação em nível de característica usando o SAPair no domínio de aparelhos celulares

(ii) Para o domínio de filmes

Módulo Indexador			
	Classe 1 (+)	Classe 2 (-)	Classe 3 (0)
Precisão	85,94%	80,00%	0,00%
Cobertura	35,95%	21,43%	0,00%
F-Measure	50,69%	33,80%	0,00%

Tabela 3.8 Desempenho da classificação em nível de característica usando o Módulo Indexador no domínio de filmes

SAPair			
	Classe 1 (+)	Classe 2 (-)	Classe 3 (0)
Precisão	81,0%	58,1%	0,0%
Cobertura	87,7%	45,5%	0,0%
F-Measure	84,2%	51,0%	0,0%

Tabela 3.9 Desempenho da classificação em nível de característica usando o SAPair no domínio de filmes

Após a análise das tabelas, observa-se que, analogamente às matrizes confusão com erros e acertos, a classe 3 (de pares com classificação neutra) teve

todos valores iguais a 0 pois nenhum comentário foi classificado manualmente como neutro.

Em comparação com o SAPair, a classificação utilizando o módulo indexador conseguiu se equivaler nas medidas de precisão das classes 1 dos dois domínios, com uma discreta melhora no domínio de filmes. Também pode ser considerada discreta a diferença na precisão da classe 2 para aparelhos celulares, com vantagem para o SAPair. Por outro lado, o módulo indexador teve melhor desempenho na precisão da classe 2 do domínio de filmes, mas não correspondeu às expectativas nas medidas de cobertura de ambas os domínios.

Além da grande diferença no número de documentos indexados entre o módulo proposto e o Google, que já foi comentada, outro fator que pode ter afetado o desempenho da classificação através das consultas ao indexador é o não-tratamento prévio dos comentários, o que permitiu que houvesse casos de comentários repetidos, escritos com erros na grafia das palavras e/ou gírias. Estes erros não aconteciam no SAPair devido ao tratamento que era feito no passo de classificação de opiniões, ao final do módulo de classificação de sentimentos do sistema.

4. Conclusão

Este trabalho de graduação teve como objetivo implementar um módulo de indexação de conteúdos opinativos, para ser utilizado na fase de Filtro de Polaridade do método PairClassif, para que as consultas de validação de polaridade fossem feitas a um índice apenas com conteúdo opinativo ao invés de se consultar o Google.

Os testes realizados através das consultas ao módulo de indexação não demonstraram um melhor desempenho em relação ao PairClassif pelo fato que houve uma alta taxa de pares com classificação neutra, comprometendo principalmente a medida de cobertura da classificação.

Em trabalhos futuros, existe a intenção de melhorar a qualidade das consultas dos padrões de sentença, aproveitando mais recursos da biblioteca Lucene, para garantir que os resultados das consultas apresentem uma intrínseca relação entre a característica e a palavra opinativa, bem como expandir a base de opiniões de domínio específico. Também há o desejo de fazer uma integração deste módulo ao SAPair, para que os resultados advindos desta fase sejam refinados pela fase de Classificação de Opinião.

Referências Bibliográficas

[1] **A Internet em números: veja o tamanho da rede!** Disponível em: <http://canaltech.com.br/noticia/internet/A-Internet-em-numeros-veja-o-tamanho-da-rede/>. Acesso em 20/01/2013.

[2] ALVES, Melina. **O crescimento exponencial da informação na web**. Disponível em: <http://webinsider.uol.com.br/2012/03/20/o-crescimento-exponencial-da-informacao-na-web/>. Acesso em 20/01/2013.

[3] SILVA, Nelson Rocha; LIMA, Diego; BARROS, Flávia. **SAPair: Um Processo de Análise de Sentimento no Nível de Característica**. In: 4nd International Workshop on Web and Text Intelligence (WTI'12), Curitiba. 2012.

[4] SILVA, Nelson Rocha. **BestChoice: Classificação de Sentimento em Ferramentas de Expressão de Opinião**. Trabalho de Graduação em Ciência da Computação – Centro de Informática/UFPE, Recife. (2010).

[5] LIMA, Diego. **PairExtractor: Extração de Pares Livre de Domínio para Análise de Sentimentos**. Trabalho de Graduação em Ciência da Computação – Centro de Informática/UFPE, Recife. (2011).

[6] SILVA, Nelson Rocha. **PairClassif: Um Método de Classificação da Polaridade de Pares para Análise de Sentimento em nível de Característica**. Dissertação de Mestrado em Ciência da Computação – Centro de Informática/UFPE, Recife. (2013).

[7] LIU, Bing. Sentiment Analysis and Subjectivity. **Handbook of Natural Language Processing**. Flórida – USA, 2a ed, Chapman and Hall/CRC, 2010.

[8] LIU, Bing. **Sentiment Analysis and Opinion Mining**. Morgan & Claypool Publishers, May 2012.

[9] LIU, B., NARAYNAN, R., CHOUDHARY, A. **Sentiment Analysis of Conditional Sentences**. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 1, pp. 180–189. Singapore (2009)

[10] WIKIPEDIA. **Precision and Recall**. Disponível em: <http://en.wikipedia.org/wiki/Precision_and_recall>. Acesso em 15/05/2013.

[11] ESULI, A.; SEBASTIANI, F. **Sentiwordnet: a publicly available lexical resource for opinion mining**. In: Proceedings of the 5th conference on language resources and evaluation, pp.417-422. Genoa - Italy. 2006.

[12] APACHE. **Apache POI - the Java API for Microsoft Documents**. Disponível em: < http://poi.apache.org/>. Acesso em 26/02/2013.

[13] WIKIPEDIA. **Apache Lucene**. Disponível em: < http://pt.wikipedia.org/wiki/Apache_Lucene>. Acesso em 01/03/2013.

[14] APACHE. **Lucene 3.6.2 Core API**. Disponível em < http://lucene.apache.org/core/3_6_2/api/core/index.html>. Acesso em 01/03/2013.